

# Compensating for population sampling in simulations of epidemic spread on temporal contact networks

Mathieu Génois,<sup>1</sup> Christian L. Vestergaard,<sup>1</sup> Ciro Cattuto,<sup>2</sup> and Alain Barrat<sup>1,2</sup>

<sup>1</sup>*Aix Marseille Université, Université de Toulon,  
CNRS, CPT, UMR 7332, 13288 Marseille, France*

<sup>2</sup>*Data Science Laboratory, ISI Foundation, Torino, Italy*  
(Dated: November 19, 2015)

Data describing human interactions often suffer from incomplete sampling of the underlying population. As a consequence, the study of contagion processes using data-driven models can lead to a severe underestimation of the epidemic risk. Here we present a systematic method to alleviate this issue and obtain a better estimation of the risk in the context of epidemic models informed by high-resolution time-resolved contact data. We consider several such data sets collected in various contexts and perform controlled resampling experiments. We show how the statistical information contained in the resampled data can be used to build a series of surrogate versions of the unknown contacts. We simulate epidemic processes on the resulting reconstructed data sets and show that it is possible to obtain good estimates of the outcome of simulations performed using the complete data set. We discuss limitations and potential improvements of our method.

Human interactions play an important role in determining the potential transmission routes of infectious diseases and other contagion phenomena [1]. Their measure and characterisation thus represent an invaluable contribution to the study of transmissible diseases [2]. While surveys and diaries in which volunteer participants record their encounters [3–7] have provided crucial insights (see however [4, 8, 9] for recent investigations of the memory biases inherent in self-reporting procedures), new approaches have recently emerged to measure contact patterns between individuals with high resolution, using wearable sensors that can detect the proximity of other similar devices [10–20]. The resulting measuring infrastructures register contacts specifically within the closed population formed by the participants wearing sensors, with typically high spatial and temporal resolutions. In the recent years, several data gathering efforts have used such methods to obtain, analyse and publish data sets describing the contact patterns between individuals in various contexts in the form of temporal networks [14, 20–24]: nodes represent individuals and, at each time step, a link is drawn between pairs of individuals who are in contact [25]. Such data has been used to inform models of epidemic spreading phenomena used to evaluate epidemic risks and mitigation strategies in specific, size-limited contexts such as schools or hospitals [14, 19, 20, 22, 26–32], finding in particular outcomes consistent with observed outbreak data [20] or providing evidence of links between specific contacts and transmission events [19, 31].

Despite the relevance and interest of such detailed data sets, as illustrated by these recent investigations, they suffer from the intrinsic limitation of the data gathering method: contacts are registered only between participants wearing sensors. Contacts with and between individuals who do not wear sensors are thus missed. In other words, as most often not all individuals accept to participate by wearing sensors, many data sets obtained by such techniques suffer from population sampling, de-

spite efforts to maximise participation through e.g. scientific engagement of participants [24, 33]. Hence, the collected data only contains information on contacts occurring among a fraction of the population under study.

Population sampling is well-known to affect the properties of static networks [34–36]: various statistical properties and mixing patterns of the contact network of a fraction of the population of interest may differ from those of the whole population, even if the sampling is uniform [37–40], and several works have focused on inferring network statistics from the knowledge of incomplete network data [39, 41–44]. Both structural and temporal properties of time-varying networks might as well be affected by missing data effects [16, 39].

In addition, a crucial though little studied consequence of such missing data is that simulations of dynamical processes in data-driven models can be affected if incomplete data are used [38, 39, 45]. For instance, in simulations of epidemic spreading, excluded nodes are by definition unreachable and thus equivalent to immunised nodes. Due to herd vaccination effects, the outcome of simulations of epidemic models on sampled networks is thus expected to be underestimated with respect to simulations on the whole network. (We note however, that in the different context of transportation networks, it was found in [45] that the inclusion of the most important transportation nodes can be sufficient to describe the global worldwide spread of influenza-like illnesses, at least in terms of times of arrival of the spread in various cities.) How to estimate the outcome of dynamical processes on contact networks using incomplete data remains an open question.

Here we make progresses on this issue for incompletely sampled data describing networks of human face-to-face interactions, collected by infrastructures based on sensors, under the assumption that the population participating to the data collection is a uniform random sample of the whole population of interest. (We do not therefore address here the issue of non-uniform sampling of contacts that may result from other measurement methods

such as diaries or surveys.) We proceed through resampling experiments on empirical data sets in which we exclude uniformly at random a fraction of the individuals (nodes of the contact network). We measure how relevant network statistics vary under such uniform resampling and confirm that, although some crucial properties are stable, numerical simulations of spreading processes performed using incomplete data lead to strong underestimations of the epidemic risk. Our goal and main contribution consists then in putting forward and comparing a hierarchy of systematic methods to provide better estimates of the outcome of models of epidemic spread in the whole population under study. To this aim, we do not try to infer the true sequence of missing contacts. Instead, the methods we present consist in the construction of surrogate contact sequences for the excluded nodes, using only structural and temporal information available in the resampled contact data. We perform simulations of spreading processes on the reconstructed data sets, obtained by the union of the resampled and surrogate contacts, and investigate how their outcomes vary depending on the amount of information included in the reconstruction method. We show that it is possible to obtain outcomes close to the results obtained on the complete data set, while, as mentioned above, using only the incomplete data severely underestimates the epidemic risk. We show the efficiency of our procedure using three data sets collected in widely different contexts and representative of very different population structures found in day-to-day life: a scientific conference, a high school and a workplace. We finally discuss the limitations of our method in terms of sampling range, model parameters and population sizes.

## RESULTS

### Data and Methodology

We consider data sets describing contacts between individuals, collected by the SocioPatterns collaboration (<http://www.sociopatterns.org>) in three different settings: a workplace (office building, InVS) [46], a high school (Thiers13) [24] and a scientific conference (SFHH) [21, 22]. These data correspond to the close face-to-face proximity of individuals equipped with wearable sensors, at a temporal resolution of 20 seconds [16]. Table I summarises the characteristics of each data set. The contact data are represented by temporal networks, in which nodes represent the participating individuals and a link between two nodes  $i$  and  $j$  at time  $t$  indicates that the two corresponding persons were in contact at that time. These three data sets were chosen as representative of different types of day-to-day contexts and of different contact network structures: the SFHH data correspond to a rather homogeneous contact network; the InVS and Thiers13 populations were instead structured in departments and classes, respectively. Moreover, high school

classes (Thiers13) are of similar sizes while the InVS department sizes are unequal. Finally, the high school contact patterns (Thiers13) are constrained by strict and repetitive school schedules, while contacts in offices are less regular across days.

To quantify how the incompleteness of data, assumed to stem from a uniformly random participation of individuals to the data collection, affects the outcome of simulations of dynamical processes, we consider as ground truth the available data and perform population resampling experiments by removing a fraction  $f$  of the nodes uniformly at random. (Note that the full data sets are also samples of all the contacts that occurred in the populations, as the participation rate was lower than 100% in each case. In the Thiers13 case however, the participation rate was quite high.) We then simulate on the resampled data the paradigmatic Susceptible-Infectious-Recovered (SIR) and the Susceptible-Infectious-Susceptible (SIS) models of epidemic propagation. In these models, a susceptible (S) node becomes infectious (I) at rate  $\beta$  when in contact with an infectious node. Infectious nodes recover spontaneously at rate  $\mu$ . In the SIR model, nodes then enter an immune recovered (R) state, while in the SIS model, nodes become susceptible again and can be reinfected. The quantities of interest are for the SIR model the distribution of epidemic sizes, defined as the final fraction of recovered nodes, and for the SIS model the average fraction of infectious nodes  $i_\infty$  in the stationary state. We also calculate for the SIR model the fraction of epidemics that infect more than 20% of the population and the average size of these epidemics. For the SIS model, we determine the epidemic threshold  $\beta_c$  for different values of  $\mu$ : it corresponds to the value of  $\beta$  that separates an epidemic-free state ( $i_\infty = 0$ ) for  $\beta < \beta_c$  from an endemic state ( $i_\infty > 0$ ) for  $\beta > \beta_c$ , and is thus an important indicator of the epidemic risk. We refer to the Methods section for further details on the simulations.

We then present several methods for constructing surrogate data using only information contained in the resampled data. We compare for each data set the outcomes of simulations performed on the whole data set, on resampled data sets with a varying fraction of nodes removed,  $f$ , and on the reconstructed data sets built using these various methods.

### Uniformly resampled contact networks

Missing data are known to affect the various properties of contact networks in different ways. In particular, the number of neighbours (degree) of a node decreases as the fraction  $f$  of removed nodes increases, since removing nodes also removes links to these nodes. Under the hypothesis of uniform sampling, the average degree  $\langle k \rangle$  becomes  $(1 - f)\langle k \rangle$  for the resampled network [47]. As a result, the density of the resampled aggregated contact network, defined as the number of links divided by

the total number of possible links between the nodes, does not depend on  $f$ . The same reasoning applies to the density  $\rho_{AB}$  of links between groups of nodes  $A$  and  $B$ , defined as the number of links  $E_{AB}$  between nodes of group  $A$  and nodes of group  $B$ , normalised by the maximum possible number of such links,  $n_A n_B$ , where  $n_A$  is the number of nodes of group  $A$  (for  $A = B$ , the maximum possible number of links is  $n_A(n_A - 1)/2$ ): both the expected number of neighbours of group  $B$  for nodes of group  $A$  (given by  $E_{AB}/n_A$ ) and the number  $n_B$  of nodes in group  $B$  are indeed reduced by a factor  $(1 - f)$ , so that  $\rho_{AB}$  remains constant. This means that the link density contact matrix, which gathers these densities and gives a measure of the interaction between groups (here classes or departments), is stable under uniform resampling. We illustrate these results on our empirical data sets in supplementary figures 1, 2, 4 and 5. Table II and supplementary figure 2 show in particular that the similarities between the original and resampled matrices are high for all data sets (see supplementary figures 4–5 for the contact matrices themselves).

Finally, the temporal statistics of the contact network are not affected by population sampling, as already noted in [16] for other data sets: the distributions of contact and inter-contact durations (the inter-contact durations are the times between consecutive contacts on a link), of number of contacts per link and of cumulated contact durations (i.e., of the link weights in the aggregated network) do not change when the network is sampled uniformly (supplementary figure 1). In the case of structured population, an interesting property is moreover illustrated in supplementary figures 6–7: although the distributions of contact durations occurring between members of the same group or between individuals belonging to different groups are indistinguishable, this is not the case for the distributions of the numbers of contacts per link nor, as a consequence, for the distributions of cumulated contact durations. In fact, both cumulated contact durations and numbers of contacts per link are more broadly distributed for links joining members of the same group. The figures show that this property is stable under uniform resampling.

Despite the robustness of these properties, the outcome of simulations of epidemic spread is strongly affected by the resampling. As Fig. 1 illustrates for instance, the probability of large outbreaks in the SIR model decreases strongly as  $f$  increases and even vanishes at large  $f$ . As mentioned above, such a result is expected, since the removed nodes act as if they were immunised: sampling hinders the propagation in simulations by removing transmission routes between the remaining nodes. As a consequence, the prevalence and the final size of the outbreaks are systematically underestimated by simulations of the SIR model on the resampled network with respect to simulations on the whole data set (for the SIS model, the epidemic threshold is overestimated): resampling leads overall to a systematic underestimation of the epidemic risk, and Fig. 1 illustrates the extent of this un-

derestimation for the data at hand.

### Estimation of epidemic sizes through simulations on reconstructed temporal networks

We now present a series of methods to improve the estimation of the epidemic risk in simulations of epidemic spread on temporal network data sets in which nodes (individuals) are missing uniformly at random. Note that we do not address here the problem of link prediction [48] as our aim is not to infer the missing contacts. The hierarchy of methods we put forward uses increasing amounts of information corresponding to increasing amounts of detail on the group and temporal structure of the contact patterns, as measured in the resampled network. We moreover assume that the timelines of scheduled activity are known (i.e., nights and weekends, during which no contact occurs).

For each data set, considered as ground truth, we create resampled data sets by removing at random a fraction  $f$  of the  $N$  nodes. We then measure on each resampled data set a series of statistics of the resulting contact network and construct stochastic, surrogate versions of the missing part of the network by creating for each missing node a surrogate instance of its links and a synthetic timeline of contacts on each surrogate link, in the different ways described below (see Supplementary Information and Methods section for more details on their practical implementation).

**Method 0.** As discussed above, the first effect of missing data is to decrease the average degree of the aggregate contact network, while keeping its density constant. Hence, the simplest approach is to merely compensate this decrease. We therefore measure the density of the resampled contact network  $\rho_s$ , as well as the average aggregate duration of the contacts,  $\langle w \rangle_s$ . We then add back the missing nodes and create surrogate links between these nodes and between these nodes and the nodes of the resampled data set at random, with the only constraint to keep the overall link density fixed to  $\rho_s$ . We then attribute to each surrogate link the same weight  $\langle w \rangle_s$  and create for each link a timeline of randomly chosen contact events of equal length  $\Delta t = 20s$  (the temporal resolution of the data set) whose total duration gives back  $\langle w \rangle_s$ .

**Method W.** The heterogeneity of aggregated contact durations is known to play a role in the spreading patterns of model diseases [4, 20, 22, 49]. We therefore refine Method 0 by collecting in the resampled data the list  $\{w\}$  of aggregate contact durations, or weights (W). We build the surrogate links and surrogate timelines of contacts on each link as in Method 0, except that each surrogate link carries a weight extracted at random from  $\{w\}$ , instead of the average  $\langle w \rangle_s$ .

**Method WS.** The fact that the population is divided into groups of individuals such as classes or departments can have a strong impact on the structure of the contact network [20, 23] and on spreading processes [50]. We

thus measure here the link density contact matrix of the resampled data, and construct surrogate links in a way to keep this matrix fixed (equal to the value measured in the resampled data), in the spirit of stochastic block models with fixed numbers of edges between blocks [51]. Moreover, we collect in the resampled data two separate lists of aggregate contact durations:  $\{w\}^{\text{int}}$  gathers the weights of links between individuals belonging to the same group, and  $\{w\}^{\text{ext}}$  is built with the weights of links joining individuals of different groups. For each surrogate link, its weight is extracted at random either from  $\{w\}^{\text{int}}$  if it joins individuals of the same group or from  $\{w\}^{\text{ext}}$  if it associates individuals of different groups. Timelines are then attributed to links as in W. This method assumes that the number of missing nodes in each group is known, and preserves the group structure (S) of the population.

**Method WT.** Several works have investigated how the temporal characteristics of networks (such as burstiness) can slow down or accelerate spreading [25, 29, 52]. In order to take these characteristics into account, we measure in the resampled data the distributions of number of contacts per link and of contact and inter-contact durations, in addition to the global network density. We build surrogate links as in Method W, and construct on each link a synthetic timeline in a way to respect the measured temporal statistics (T) of contacts. More precisely, we attribute at random a number of contacts (taken from the measured distribution) to each surrogate link, and then alternate contact and inter-contact durations taken at random from the respective empirical distributions.

**Method WST.** This method conserves the distribution of link weights (W), the group structure (S), and the temporal characteristics of contacts (T): surrogate links are built and weights assigned as in method WS, and contact timelines on each link as in method WT.

Each of these methods uses a different amount of information gathered from the resampled data. Methods 0, W and WT include an increasing amount of detail on the temporal structure of contacts: method 0 assumes homogeneity of aggregated contact durations, while W takes into account their heterogeneity, and WT reproduces heterogeneities of contact and inter-contact durations. On the other hand, neither of these three methods assume any knowledge of the population group structure. This can be due either to an effective complete lack of knowledge about the population structure, as in the SFHH data, or also to the lack of data on the repartition of the missing nodes in the groups. Methods WS and WST on the other hand reproduce the group structure as in a stochastic block model with fixed number of links within and between groups, and take into account the difference between the distributions of numbers of contacts and aggregate durations between individuals of the same or of different groups. Indeed, links within groups correspond on average to larger weights, as found empirically in [50] and discussed above (supplementary figures 5–6). Overall, method WST is the one that uses most information

measured in the resampled data. (Additional properties such as the transitivity -which is also stable under resampling procedure, see supplementary figure 3- can also be measured in the resampled data and imposed in the construction of surrogate links, as detailed in the Supplementary Information. This comes however at a strong computational cost and we have verified that it does not impact significantly our results, as shown in the supplementary figure 20.)

We check in Table II and supplementary figures 8–13 that the statistical properties of the resulting reconstructed (surrogate) networks, obtained by the union of the resampled data and of the surrogate links, are similar to the ones of the original data for the WST method. We emphasise again that our aim is not to infer the true missing contacts, so that we do not compare the detailed structures of the surrogate and original contact networks.

Figures 2, 3, 4 and supplementary figures 16–19 display the outcome of SIR spreading simulations performed on surrogate networks obtained using the various reconstruction methods, compared with the outcome of simulations on the resampled data sets, for various values of  $f$ . Method 0 leads to a clear overestimation of the outcome and does not capture well the shape of the distribution of outbreak sizes. Method W gives only slightly better results. The overall shape of the distribution is better captured for the three reconstruction methods using more information: WS, WT and WST (note that for the SFHH case the population is not structured, so that W and WS are equivalent, as are WT and WST). The WST method matches best the shape of the distributions and yields distributions much more similar to those obtained by simulating on the whole data set than the simulations performed on the resampled networks. We also show in Fig. 5 the fraction of outbreaks that reach at least 20% of the population and the average epidemic size for these outbreaks. In the case of simulations performed on resampled data, we rapidly lose information about the size and even the existence of large outbreaks as  $f$  increases. Simulations using data reconstructed with methods 0 and W, on the contrary, largely overestimate these quantities, which is expected as infections spread easier on random graphs than on structured graphs [50, 52], especially if the heterogeneity of the aggregated contact durations is not considered [20, 22]. Taking into account the population structure or using contact sequences that respect the temporal heterogeneities (broad distributions of contact and inter-contact durations) yield better results (WS and WT cases, respectively). Overall, the WST method, for which the surrogate networks respect all these constraints, yields the best results.

We show in the Supplementary Information that similar results are obtained for different values of the spreading parameters. Moreover, as shown in Fig. 6 and supplementary figures 14–15, the phase diagram obtained for the SIS model when using reconstructed networks is much closer to the original than for resampled networks. Overall, simulations on networks reconstructed

using the WST method yield a much better estimation of the epidemic risk than simulations using resampled network data, for both SIS and SIR models.

### Reshuffled data sets.

Even when simulations are performed on reconstructed contact patterns built with the WST method, the maximal outbreak sizes are systematically overestimated (Figs. 2 - 4), as well as, in most cases, the probability and average size of large outbreaks, especially for the SFHH case (Figs. 4 - 5). These discrepancies might stem from structural and/or temporal correlations present in empirical contact data that are not taken into account in our reconstruction methods. In order to test this hypothesis, we construct several reshuffled data sets and use them as initial data in our resampling and reconstruction procedure. We use both structural and temporal reshuffling as described in the Methods section, in order to remove either structural correlations, temporal correlations, or both, from the original data sets. We then proceed to a resampling and reconstruction procedure (using the WST method) as for the original data, and perform numerical simulations of SIR processes. As for the original data, simulations on resampled data lead to a strong underestimation of the process outcome, and simulations using the reconstructed data gives much better results.

We show in the supplementary figures 21-22 that we still obtain discrepancies, and in particular overestimations of the largest epidemic sizes, when we use temporally reshuffled data in which the link structure of the contact network is maintained. If on the other hand we use data in which the network structure has been reshuffled in a way to cancel structural correlations within each group, the reconstruction procedure gives a very good agreement between the distributions of epidemic sizes of original and reconstructed data, as shown in Fig. 7. More precisely we consider here “CM-shuffled” data, i.e., contact networks in which the links have been reshuffled randomly but separately for each pair of groups, i.e., a link between an individual of group  $A$  and an individual of group  $B$  is still between groups  $A$  and  $B$  in the reshuffled network. The difference with the case of non-reshuffled empirical data is particularly clear for the SFHH case. This indicates that the overestimation observed in Figs. 2 - 4 is mostly due to the fact that the reconstructed data does not reproduce small scale structures of the contact networks: such structures might be due to e.g. groups of colleagues or friends, whose composition is neither available as metadata nor detectable in the resampled data sets.

### Limitations.

When the fraction  $f$  of nodes excluded by the resampling procedure becomes large, the properties of the re-

sampled data may start to differ substantially from those of the whole data set (Figs. S1 & S2). As a result, the distributions of epidemic sizes of SIR simulations show stronger deviations from those obtained on the whole data set (Fig. 8), even if the epidemic risk evaluation is still better than for simulations on the resampled networks (Fig. 5). Most importantly however, the information remaining in the resampled data at large  $f$  can be insufficient to construct surrogate contacts. This happens in particular if an entire class or department is absent from the resampled data or if all the resampled nodes of a class/department are disconnected (see Methods for details). We show in the bottom plots of Fig. 5 the failure rate, i.e., the fraction of cases in which we are not able to construct surrogate networks from the resampled data. The failure rate increases gradually with  $f$  for the InVS data since the groups (departments) are of different sizes. For the Thiers13 data, all classes are of similar sizes so that the failure rate reaches abruptly a large value at a given value of  $f$ . For the SFHH data, we can always construct surrogate networks as the population is not structured. Another limitation of the reconstruction method lies in the need to know the number of individuals missing in each department or class. If these numbers are completely unknown, giving an estimation of outbreak sizes is impossible as adding arbitrary numbers of nodes and links to the resampled data can lead to arbitrarily large epidemics. The methods are however still usable if only partial information is available. For instance, if only the overall missing number of individuals is available, it is possible to use the WT method, which still gives sensible results. Moreover, if  $f$  is only approximately known, e.g.,  $f$  is known to be within an interval of possible values  $[f_1, f_2]$ , it is possible to perform two reconstructions using the respective hypothesis  $f = f_1$  and  $f = f_2$  and to give an interval of estimates. We provide an example of such procedure in supplementary figure 23.

## DISCUSSION

The understanding of epidemic spreading phenomena has been vastly improved thanks to the use of data-driven models at different scales. High resolution contact data in particular have been used to evaluate epidemic risk or containment policies in specific populations or to perform contact tracing [14, 19, 20, 28, 30–32]. In such studies, missing data due to population sampling might represent however a serious issue: individuals absent from a data set are indeed equivalent to immunised individuals when epidemic processes are simulated. Feeding sampled data into data-driven models can therefore lead to severe underestimations of the epidemic risk and might even a priori affect the evaluation of mitigation strategies if for instance some at-risk groups are particularly undersampled.

Here we have put forward a set of methods to obtain a better evaluation of the outcome of spreading simu-

lations for data-driven models using contact data from a uniformly sampled population. To this aim, we have shown how it is possible, starting from a data set describing the contacts of only a fraction of the population of interest (uniformly sampled from the whole population), to construct surrogate data sets using various amounts of accessible information, i.e., quantities measured in the sampled data. We have shown that the simplest method, which consists in simply compensating for the decrease in the average number of neighbours due to sampling, yields a strong overestimation of the epidemic risk. When additional information describing the group structure and the temporal properties of the data is added in the construction of surrogate data sets, simulations of epidemic spreading on such surrogate data yield results similar to those obtained on the complete data set. (We note that the issue of how much information should be included when constructing the surrogate data is linked to the general issue of how much information is needed to get an accurate picture of spreading processes on temporal networks [22, 27, 28, 32, 53, 54].) Some discrepancies in the epidemic risk estimation are however still observed, due in particular to small scale structural correlations of the contact network that are difficult or even impossible to measure in the resampled data: these discrepancies are indeed largely suppressed if we use as original data a reshuffled contact network in which such correlations are absent.

The methods presented here yield much better results than simulations using resampled data, even when a substantial part of the population is excluded, in particular in estimating the probability of large outbreaks. It suffers however from limitations, especially when the fraction  $f$  of excluded individuals is too large. First, the construction of the surrogate contacts relies on the stability of a set of quantities with respect to resampling, but the measured quantities start to deviate from the original ones at large  $f$ . The shape of the distribution of epidemic sizes may then differ substantially from the original one. Second, large values of  $f$  might even render the construction of the surrogate data impossible due to the loss of information on whole categories of nodes. Finally, at least an estimate of the number of missing individuals in the population is needed in order to create a surrogate data set.

An interesting avenue for future work concerns possible improvements of the reconstruction methods, in particular by integrating into the surrogate data additional information and complex correlation patterns measured in the sampled data. For instance, the number of contacts varies significantly with the time of day in most contexts: the corresponding activity timeline might be measured in the sampled data (overall or even for each group of individuals), assumed to be robust to sampling and used in the reconstruction of contact timelines. More systematically, it might also be possible to use the temporal network decomposition technique put forward in [55] on the sampled data, in order to extract mesostructures such

as temporally-localized mixing patterns. The surrogate contacts could then be built in a way to preserve such patterns. Indeed, correlations between structure and activity in the temporal contact network are known to influence spreading processes [21, 52, 54, 56–58] but are notoriously difficult to measure. If the group structure of the population is unknown, recent approaches based on stochastic block models [59] might be used to extract groups from the resampled data; this extracted group structure could then be used to build the corresponding contact matrix and surrogate data sets.

We finally recall that we have assumed an uniform sampling of nodes, corresponding to an independent random choice of each individual of the population to take part or not to the data collection. Other types of sampling or data losses can however also be present in data collected by wearable sensors, such as partial coverage of the premises of interest by the measuring infrastructure, non-uniform sampling depending on individual activity (too busy persons or, on the contrary, asocial individuals, might not want to wear sensors), on group membership, or due to clusters of non-participating individuals (e.g., groups of friends). In addition, other types of data sets such as the ones obtained from surveys or diaries correspond to different types of sampling, as each respondent provides then information in the form of an ego-network [60]. Such data sets involve potentially additional types of biases such as underreporting of the number of contacts and overestimation of contact durations [8, 9, 61]: how to adapt the methods presented here is an important issue that we will examine in future work. Finally, the population under study is (usually) not isolated from the external world, and it would be important to devise ways to include contacts with outsiders in the data and simulations, for instance by using other data sources such as surveys.

## ACKNOWLEDGEMENTS

The present work is partially supported by the French ANR project HarMS-flu (ANR-12-MONU-0018) to M.G. and A.B., by the EU FET project Multiplex 317532 to A.B., C.C. and C.L.V., by the A\*MIDEX project (ANR-11-IDEX-0001-02) funded by the “Investissements d’Avenir” French Government program, managed by the French National Research Agency (ANR) to A.B., by the Lagrange Project of the ISI Foundation funded by the CRT Foundation to C.C., and by the Q-ARACNE project funded by the Fondazione Compagnia di San Paolo to C.C.

## AUTHOR CONTRIBUTIONS

A.B. and C.C. designed and supervised the study. M.G., C.L.V., C.C., and A.B. collected and post-processed the data, analyzed the data, carried out com-

puter simulations and prepared the figures. M.G., C.L.V., C.C., and A.B. wrote the manuscript.

The authors declare that no competing financial interests exist.

## METHODS

### Data

We consider data sets collected using the SocioPatterns proximity sensing platform (<http://www.sociopatterns.org>) based on wearable sensors that detect close face-to-face proximity of individuals wearing them. Informed consent was obtained from all participants and the French national bodies responsible for ethics and privacy, the Commission Nationale de l'Informatique et des Libertés (CNIL, <http://www.cnil.fr>), approved the data collections.

The high school (Thiers13) data set [61] is structured in 9 classes, forming three subgroups of three classes corresponding to their specialisation in Mathematics-Physics (MP, MP\*1, MP\*2 with respectively 31, 29 and 38 students), Physics (PC, PC\*, PSI with respectively 44, 39 and 34 students), or Biology (2BIO1, 2BIO2, 2BIO3 with respectively 37, 35 and 39 students).

The workplace (InVS) data set [46] is structured in 5 departments: DISQ (Scientific Direction, 15 persons), DMCT (Department of Chronic Diseases and Traumatism, 26 persons), DSE (Department of Health and Environment, 34 persons), SRH (Human Resources, 13 persons) and SFLE (Logistics, 4 persons).

For the conference data (SFHH), we do not have meta-data on the participants, and the aggregated network structure was found to be homogeneous [22].

### SIR and SIS simulations

Simulations of SIR and SIS processes on the temporal networks of contacts (original, resampled or reconstructed) are performed using the temporal Gillespie algorithm described in [62]. For each run of the simulations, all nodes are initially susceptible; a node is chosen at random as the seed of the epidemic and put in the infectious state at a point in time chosen at random over the duration of the contact data. A susceptible node in contact with an infectious node becomes infectious at rate  $\beta$ . Infectious nodes recover at rate  $\mu$ : in the SIR model they then enter the recovered state and cannot become infectious again, while in the SIS model they enter the susceptible state again. If needed, the sequence of contacts is repeated in the simulation [22].

For SIR processes, we run each simulation, with the seed node chosen at random, until no infectious individual remains (nodes are thus either still susceptible or have been infected and then recovered). We consider values of  $\beta$  and  $\mu$  yielding a non-negligible epidemic risk, i.e., such

that a rather large fraction of simulations lead to a final size larger than 20% of the population (see Figs. 1–4):  $\beta = 4 \times 10^{-4} s^{-1}$ ,  $\mu = 4 \times 10^{-7} s^{-1}$  (InVS) or  $4 \times 10^{-6} s^{-1}$  (SFHH and Thiers13). Other parameter values are explored in the Supplementary Information. For each set of parameters, the distribution of epidemic sizes is obtained by performing 1,000 simulations.

For SIS processes, simulations are performed using the quasi-stationary approach of [63]. They are run until the system enters a stationary state as witnessed by the mean number of infected nodes being constant over time. Simulations are then continued for 50,000 time-steps while recording the number of infected nodes. For each set of parameters, the simulations are performed once with each node of the network chosen as the seed node.

### Reconstruction algorithm

We consider a population  $\mathcal{P}$  of  $N$  individuals, potentially organised in groups. We assume that all the contacts occurring among a subpopulation  $\tilde{\mathcal{P}}$  of these individuals, of size  $\tilde{N} = (1 - f)N$ , are known. This constitutes our resampled data from which we need to construct a surrogate set of contacts concerning the remaining  $n = N - \tilde{N} = fN$  individuals for which no contact information is available: these contacts can occur among these individuals and between them and the members of  $\tilde{\mathcal{P}}$ . We assume that we know the group of each member of  $\mathcal{P} \setminus \tilde{\mathcal{P}}$ , and the overall activity timeline, i.e., the intervals during which contacts take place, separated by nights and weekends.

To construct the surrogate data (WST method), we first compute from the activity timeline the total duration  $T_u$  of the periods during which contacts can occur.

Then, we measure in the sampled data:

- the density  $\rho$  of links in the aggregated contact network;
- a row-normalised contact matrix  $C$ , in which the element  $C_{AB}$  gives the probability for a node of group  $A$  to have a link to a node of group  $B$ ;
- the list  $\{\tau_c\}$  of contact durations;
- the lists  $\{\tau_{ic}\}^{\text{int}}$  and  $\{\tau_{ic}\}^{\text{ext}}$  of inter-contact durations for internal and external links, i.e., for links between nodes of the same group and links between nodes that belong to two different groups, respectively;
- the lists  $\{p\}^{\text{int}}$  and  $\{p\}^{\text{ext}}$  of numbers of contacts per link, respectively for internal (within groups) and external (between groups) links;
- the list  $\{t_0\}$  of initial times between the start of the data set and the first contact between two nodes.

Given  $\rho$ , we compute the number  $e$  of additional links needed to keep the network density constant when we add the  $n$  excluded nodes.

We then construct each link according to the following procedure:

- a node  $i$  is randomly chosen from the set  $\mathcal{P} \setminus \tilde{\mathcal{P}}$  of excluded nodes;
- knowing the group  $A$  that  $i$  belongs to, we extract at random a target group  $B$  with probability given by  $C_{AB}$ ;
- we draw a target node  $j$  at random from  $B$  (if  $B = A$ , we take care that  $i \neq j$ ) such that  $i$  and  $j$  are not linked;
- depending on whether  $i$  and  $j$  belong to the same group or not, we draw from  $\{p\}^{\text{int}}$  or  $\{p\}^{\text{ext}}$  the number of contact events  $p$  taking place over the link  $ij$ ;
- from  $\{t_0\}$ , we draw the initial waiting time before the first contact;
- from  $\{\tau_c\}$ , we draw  $p$  contact durations  $\tau_c^k$ ,  $k = 1, \dots, p$ ;
- from  $\{\tau_{ic}\}^{\text{int}}$  or  $\{\tau_{ic}\}^{\text{ext}}$ , we draw  $p-1$  inter-contact durations  $\tau_{ic}^m$ ,  $m = 1, \dots, p-1$ ;
- if  $t_0 + \sum_k \tau_c^k + \sum_m \tau_{ic}^m > T_u$ , we repeat steps (d) to (g) until we obtain a set of values such that  $t_0 + \sum_k \tau_c^k + \sum_m \tau_{ic}^m \leq T_u$ ;
- from  $t_0$  and the  $\tau_c^k$  and  $\tau_{ic}^m$ , we build the contact timeline of the link  $ij$ ;
- finally, we insert in the contact timeline the breaks defined by the global activity timeline.

#### Possible failure of the reconstruction method at large $f$

The construction of the surrogate version of the missing links uses as an input the group structure of the subgraph that remains after sampling, as given by the contact matrix of the link densities between the different groups of nodes that are present in the subpopulation  $\tilde{\mathcal{P}}$ . Depending on the characteristics of  $\tilde{\mathcal{P}}$  and of the corresponding contacts, the construction method can fail in several cases: (i) if an entire group (class/department) of nodes in the population is absent from  $\tilde{\mathcal{P}}$ ; (ii) if the remaining nodes of a specific group (class/department) are all isolated in  $\tilde{\mathcal{P}}$ 's contact network; (iii) if, during the algorithm, a node of  $\mathcal{P} \setminus \tilde{\mathcal{P}}$  is selected in a certain group  $A$  but cannot create any more links because it already has links to all nodes in the groups  $B$  such that  $C_{AB} \neq 0$ ; (iv) if there are either no internal (within groups) or external (between groups) links in the contact network of

$\tilde{\mathcal{P}}$ : in this case one of the lists of link temporal characteristics is empty and the corresponding structures cannot be reconstructed.

Cases (i) and (ii) correspond to a complete loss of information about the connectivity of a group (class/department) of the population, due to sampling. It is then impossible to reconstruct a sensible connectivity pattern for these nodes. Case (iii) is more subtle and occurs in situations of very low connectivity between groups. For instance, within the contact network of  $\mathcal{P}$ , a group  $A$  has links only with another specific group  $B$ , and both  $A$  and  $B$  are small; it is then possible that the nodes of  $(\mathcal{P} \setminus \tilde{\mathcal{P}}) \cap A$  exhaust the set of possible links to nodes of  $B$  during the reconstruction algorithm. If a node of  $(\mathcal{P} \setminus \tilde{\mathcal{P}}) \cap A$  is again chosen to create a link, such a creation is not possible and the construction fails. Case (iv) usually corresponds to situations in which the links between individuals of different groups which remain in the resampled data set correspond to pairs of individuals who have had only one contact event: in such cases,  $\{\tau_{ic}\}^{\text{ext}}$  is empty and external links with more than one contact cannot be built.

#### Shufflings

In order to test the effect of correlations in the temporal network, we use four shuffling methods, based on the ones defined in [56].

**Link shuffling.** The contact timelines associated with each link are randomly redistributed among the links. Correlations between timelines of links adjacent to a given node are destroyed, as well as correlations between weights and topology. The structure of the network is kept, as well as the global activity timeline.

**Time shuffling.** From the contact data we build the lists  $\{\tau_c\}$ ,  $\{\tau_{ic}\}$  and  $\{p\}$  of, respectively, contact durations, inter-contact durations and number of contacts per link. We also measure the list  $\{t_0\}$  of initial times between the start of the data set and the first contact between two nodes. For each link, we draw randomly a starting time  $t_0$ , a number  $p$  of contacts from  $\{p\}$ ,  $p$  contact durations from  $\{\tau_c\}$  and  $p-1$  inter-contact durations from  $\{\tau_{ic}\}$ , so that the total duration of the timeline does not exceed the total available time  $T_u$ . We then construct the contact timelines, thus destroying the temporal correlations among contacts. The structure of the network is instead kept fixed.

**CM shuffling.** We perform a link rewiring separately on each compartment of the contact matrix, *i.e.*, we randomly redistribute links with their contact timelines within each group, and within each pair of groups. We thus destroy the structural correlations inside each compartment of the contact matrix, while preserving the group structure of the network as given by the link density contact matrix and the contact matrix of total contact times between groups.

**CM-time shuffling.** We perform both a CM shuffling



and a time shuffling.

- 
- [1] Barrat, A., Barthélemy, M. & Vespignani, A. *Dynamical processes on complex networks* (Cambridge University Press (Cambridge), 2008).
  - [2] Read, J. M., Edmunds, W. J., Riley, S., Lessler, J. & Cummings, D. A. T. Close encounters of the infectious kind: methods to measure social mixing behaviour. *Epidemiology & Infection* **140**, 2117–2130 (2012).
  - [3] Edmunds, W. J., O’callaghan, C. J. & Nokes, D. J. Who mixes with whom? a method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **264**, 949–957 (1997).
  - [4] Read, J., Eames, K. & Edmunds, W. Dynamic social networks and the implications for the spread of infectious disease. *J R Soc Interface* **5**, 1001–7 (2008).
  - [5] Mossong, J. *et al.* Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med* **5**, e74 (2008).
  - [6] Danon, L., House, T., Read, J. & Keeling, M. Social encounter networks: collective properties and disease transmission. *J. R. Soc. Interface* **9**, 2826–2833 (2012).
  - [7] Danon, L., Read, J. M., House, T. A., Vernon, M. C. & Keeling, M. J. Social encounter networks: characterizing great britain. *Proceedings of the Royal Society B: Biological Sciences* **280**, 1765 (2013).
  - [8] Smieszek, T., Burri, E. U., Scherzinger, R. & Scholz, R. W. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiology & Infection* **140**, 744–752 (2012).
  - [9] Smieszek, T. *et al.* How should social mixing be measured: comparing web-based survey and sensor-based methods. *BMC Infectious Diseases* **14**, 136 (2014).
  - [10] Hui, P. *et al.* Pocket switched networks and human mobility in conference environments. In *WDTN ’05: Proc. 2005 ACM SIGCOMM workshop on Delay-tolerant networking* (ACM, New York, NY, USA, 2005).
  - [11] O’Neill, E. *et al.* Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *Ubicomp*, vol. 4206, 315–332 (2006).
  - [12] Eagle, N., Pentland, A. S. & Lazer, D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* **106**, 15274–15278 (2009).
  - [13] Vu, L., Nahrstedt, K., Retika, S. & Gupta, I. Joint bluetooth/wifi scanning framework for characterizing and leveraging people movement in university campus. In *Proceedings of the 13th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems*, MSWIM ’10, 257–265 (ACM, New York, NY, USA, 2010).
  - [14] Salathé, M. *et al.* A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* **107**, 22020–22025 (2010).
  - [15] Hashemian, M., Stanley, K. & Osgood, N. Flunet: Automated tracking of contacts during flu season. In *Proceedings of the 6th International workshop on Wireless Network Measurements*, 557–562 (2010).
  - [16] Cattuto, C. *et al.* Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE* **5**, e11596 (2010).
  - [17] Hornbeck, T. *et al.* Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *Journal of Infectious Diseases* **206**, 1549–1557 (2012).
  - [18] Stopczynski, A. *et al.* Measuring large-scale social networks with high resolution. *PLoS ONE* **9**, e95978 (2014).
  - [19] Obadia, T. *et al.* Detailed contact data and the dissemination of staphylococcus aureus in hospitals. *PLoS Comput Biol* **11**, e1004170 (2015).
  - [20] Toth, D. J. A. *et al.* The role of heterogeneity in contact timing and duration in network models of influenza spread in schools. *Journal of The Royal Society Interface* **12**, 20150279 (2015).
  - [21] Isella, L. *et al.* What’s in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* **271**, 166–180 (2011).
  - [22] Stehlé, J. *et al.* Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine* **9**, 87 (2011).
  - [23] Stehlé, J. *et al.* High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS ONE* **6**, e23176 (2011).
  - [24] Fournet, J. & Barrat, A. Contact patterns among high school students. *PLoS ONE* **9**, e107878 (2014).
  - [25] Holme, P. & Saramki, J. Temporal networks. *Physics Reports* **519**, 97 – 125 (2012).
  - [26] Lee, S., Rocha, L. E. C., Liljeros, F. & Holme, P. Exploiting temporal network structures of human interaction to effectively immunize populations. *PLoS ONE* **7**, e36439 (2012).
  - [27] Machens, A. *et al.* An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. *BMC Infectious Diseases* **13**, 185 (2013).
  - [28] Smieszek, T. & Salathé, M. A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks. *BMC MEDICINE* **11**, 35 (2013).
  - [29] Masuda, N. & Holme, P. Predicting and controlling infectious disease epidemics using temporal networks. *F1000Prime Reports* **5** (2013).
  - [30] Gemmetto, V., Barrat, A. & Cattuto, C. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Diseases* **14**, 695 (2014).
  - [31] Voirin, N. *et al.* Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital. *Infection Control & Hospital Epidemiology* **36**, 254–260 (2015).
  - [32] Chowell, G. & Viboud, C. A practical method to target individuals for outbreak detection and control. *BMC Medicine* **11**, 36 (2013).
  - [33] Conlan, A. J. K. *et al.* Measuring social networks in british primary schools through scientific engagement. *Proceedings of the Royal Society B: Biological Sciences* **278**, 1467–1475 (2011).

- [34] Granovetter, M. Network sampling: Some first steps. *American Journal of Sociology* **81**, pp. 1287–1303 (1976).
- [35] Frank, O. Sampling and estimation in large social networks. *Social Networks* **1**, 91 – 101 (1979).
- [36] Achlioptas, D., Clauset, A., Kempe, D. & Moore, C. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05, 694–703 (ACM, New York, NY, USA, 2005).
- [37] Kossinets, G. Effects of missing data in social networks. *Social Networks* **28**, 247 – 268 (2006).
- [38] Ghani, A. C., Donnelly, C. A. & Garnett, G. P. Sampling biases and missing data in explorations of sexual partner networks for the spread of sexually transmitted diseases. *Statistics in Medicine* **17**, 2079–2097 (1998).
- [39] Ghani, A. C. & Garnett, G. P. Measuring sexual partner networks for transmission of sexually transmitted diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **161**, 227–238 (1998).
- [40] Onnela, J.-P. & Christakis, N. A. Spreading paths in partially observed social networks. *Phys. Rev. E* **85**, 036106 (2012).
- [41] Viger, F., Barrat, A., Dall'Asta, L., Zhang, C.-H. & Kolaczyk, E. What is the real size of a sampled network? the case of the Internet. *Phys. Rev. E* **75**, 056111 (2007).
- [42] Bliss, C. A., Danforth, C. M. & Dodds, P. S. Estimation of global network statistics from incomplete data. *PLoS ONE* **9**, e108471 (2014).
- [43] Zhang, Y., Kolaczyk, E. D. & Spencer, B. D. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Ann. Appl. Stat.* **9**, 166–199 (2015).
- [44] Cimini, G., Squartini, T., Gabrielli, A. & Garlaschelli, D. Systemic risk analysis in reconstructed economic and financial networks. Preprint at <http://arxiv.org/abs/1411.7613> (2014).
- [45] Bobashev, G., Morris, R. J. & Goedecke, D. M. Sampling for global epidemic models and the topology of an international airport network. *PLoS ONE* **3**, e3154 (2008).
- [46] Génois, M. *et al.* Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science* **3**, 326–347 (2015).
- [47] Cohen, R., Erez, K., ben Avraham, D. & Havlin, S. Resilience of the Internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626–4628 (2000).
- [48] Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* **58**, 1019–1031 (2007).
- [49] Smieszek, T., Fiebig, L. & Scholz, R. Models of epidemics: when contact repetition and clustering should be included. *Theoretical Biology and Medical Modelling* **6**, 11 (2009).
- [50] Onnela, J.-P. *et al.* Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* **104**, 7332–7336 (2007).
- [51] Peixoto, T. P. Entropy of stochastic blockmodel ensembles. *Phys. Rev. E* **85**, 056122 (2012).
- [52] Karsai, M. *et al.* Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E* **83**, 025102 (2011).
- [53] Blower, S. & Go, M.-H. The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy? *BMC Medicine* **9**, 88 (2011).
- [54] Pfitzner, R., Scholtes, I., Garas, A., Tessone, C. J. & Schweitzer, F. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Physical Review Letters* **110**, 198701 (2013).
- [55] Gauvin, L., Panisson, A. & Cattuto, C. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PLOS ONE* **9**, e86028 (2014).
- [56] Gauvin, L., Panisson, A., Cattuto, C. & Barrat, A. Activity clocks: spreading dynamics on temporal networks of human contact. *Scientific reports* **3**, 3099 (2013).
- [57] Scholtes, I. *et al.* Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nat. Comm* **5**, 5024 (2014).
- [58] Gauvin, L., Panisson, A., Barrat, A. & Cattuto, C. Revealing latent factors of temporal networks for mesoscale intervention in epidemic spread. Preprint at <http://arxiv.org/abs/1501.02758> (2015).
- [59] Peixoto, T. P. Inferring the mesoscale structure of layered, edge-valued and time-varying networks. Preprint at <http://arxiv.org/abs/1504.02381> (2015).
- [60] Robins, G., Pattison, P. & Woolcock, J. Missing data in networks: exponential random graph ( $p^*$ ) models for networks with non-respondents. *Social Networks* **26**, 257 – 283 (2004).
- [61] Mastrandrea, R., Fournet, J., & Barrat, A. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. Preprint at <http://arxiv.org/abs/1506.03645> (2015).
- [62] Vestergaard, C. L. & Génois, M. Temporal gillespie algorithm: Fast simulation of contagion processes on time-varying networks. Preprint at <http://arxiv.org/abs/1504.01298v1> (2015).
- [63] Ferreira, S. C., Ferreira, R. S. & Pastor-Satorras, R. Quasistationary analysis of the contact process on annealed scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **83**, 066113 (2011).

## TABLES &amp; FIGURES

Data set	Type	$N$	$r$	$T$	Dates
InVS	Workplace	92	63 %	2 weeks	June 24th - July 5th 2013
Thiers13	High school	326	86 %	1 week	December 2nd - 7th 2013
SFHH	Conference	403	34 %	2 days	June 3rd - 4th 2009

TABLE I. **Data sets.** For each data set we specify the type of social situation, the number  $N$  of individuals whose contacts were measured, the corresponding participation rate  $r$ , the duration  $T$  and the dates of the data collection.

	$f$	InVS CML	Thiers13 CML
Resampled	10 %	0.996 [0.937, 0.999]	0.999 [0.998, 0.999]
	20 %	0.980 [0.889, 0.994]	0.996 [0.995, 0.997]
	40 %	0.925 [0.872, 0.983]	0.988 [0.983, 0.990]
Reconstructed	10 %	0.976 [0.846, 0.995]	0.998 [0.994, 0.999]
	20 %	0.942 [0.844, 0.984]	0.993 [0.985, 0.995]
	40 %	0.890 [0.652, 0.953]	0.977 [0.938, 0.987]

TABLE II. **Contact matrix similarities** Similarities between the original contact matrices and the contact matrices of the resampled networks (top) and of the reconstructed networks (bottom). Median and 90% confidence interval for the cosine similarity between link density contact matrices (CML) for different values of  $f$ , the fraction of nodes removed from the original data. Values were obtained from 100 independent realisations of the resampling and reconstruction procedures.

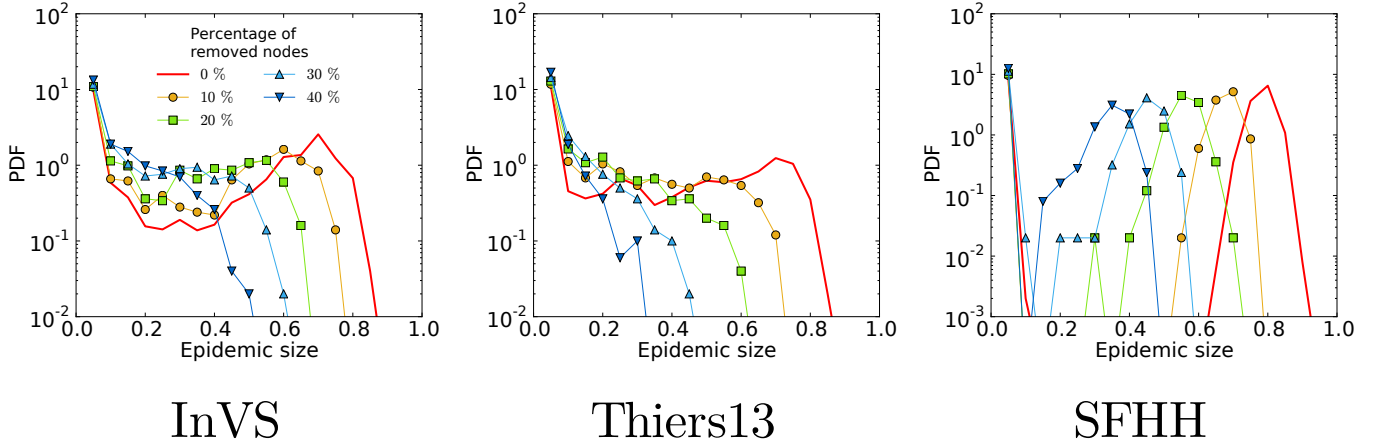


FIG. 1. **SIR epidemic simulations on resampled contact networks.** We plot the distributions of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of resampled contact networks, for different values of the fraction  $f$  of nodes removed. The plot shows the progressive disparition of large epidemic outbreaks as  $f$  increases. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (InVS) or  $\beta/\mu = 100$  (Thiers13 and SFHH). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.

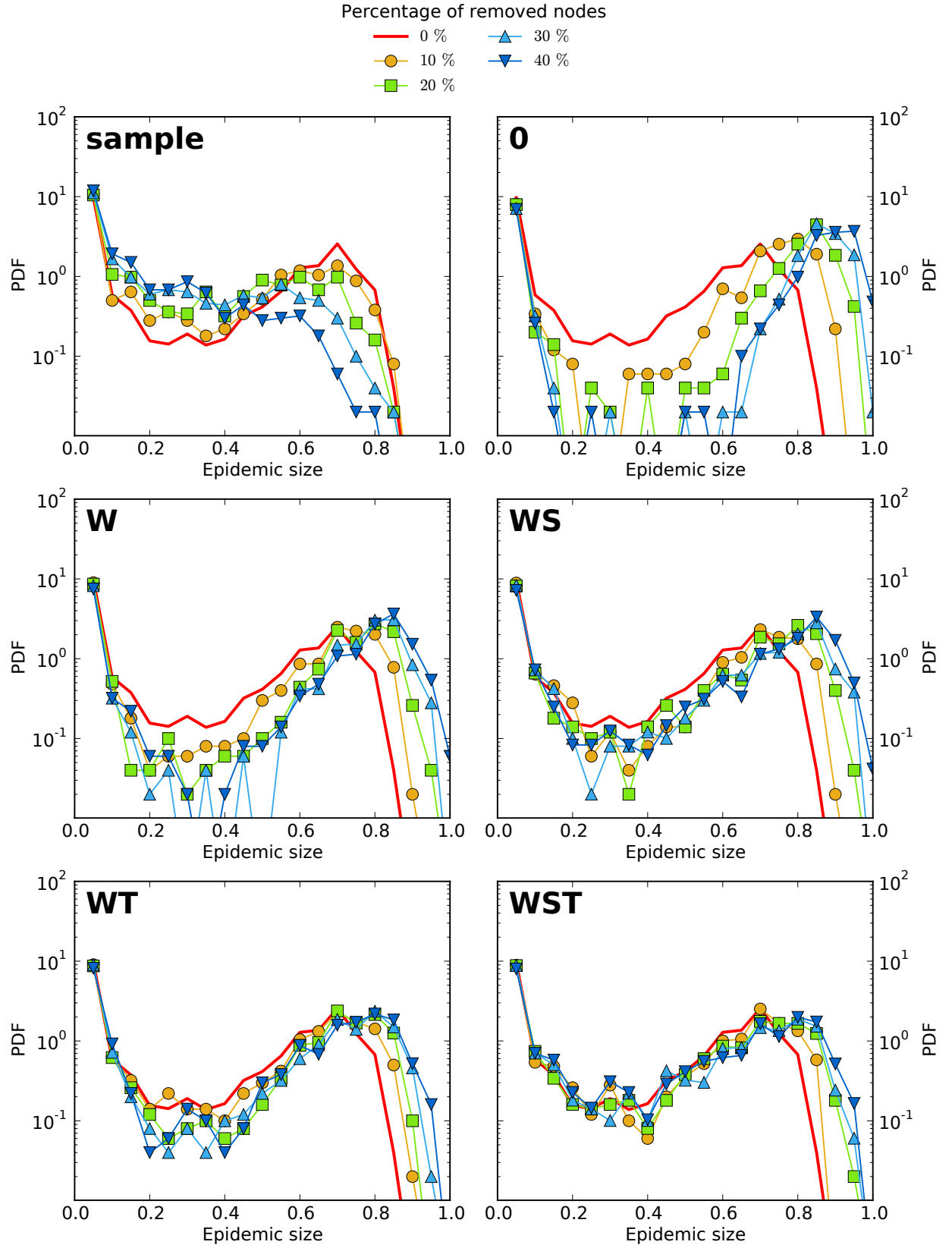


FIG. 2. **SIR simulations for the InVS (workplace) case.** We compare of the outcome of SIR epidemic simulations performed on resampled and reconstructed contact networks, for different methods of reconstruction. We plot the distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of resampled (sample) and reconstructed contact networks, for different values of the fraction  $f$  of nodes removed, and for the 5 reconstruction methods described in the text (0, W, WS, WT, WST). The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 1000$ . The case  $f = 0$  corresponds to simulations using the whole data set, *i.e.*, the reference case. For each value of  $f$ , 1,000 independent simulations were performed.

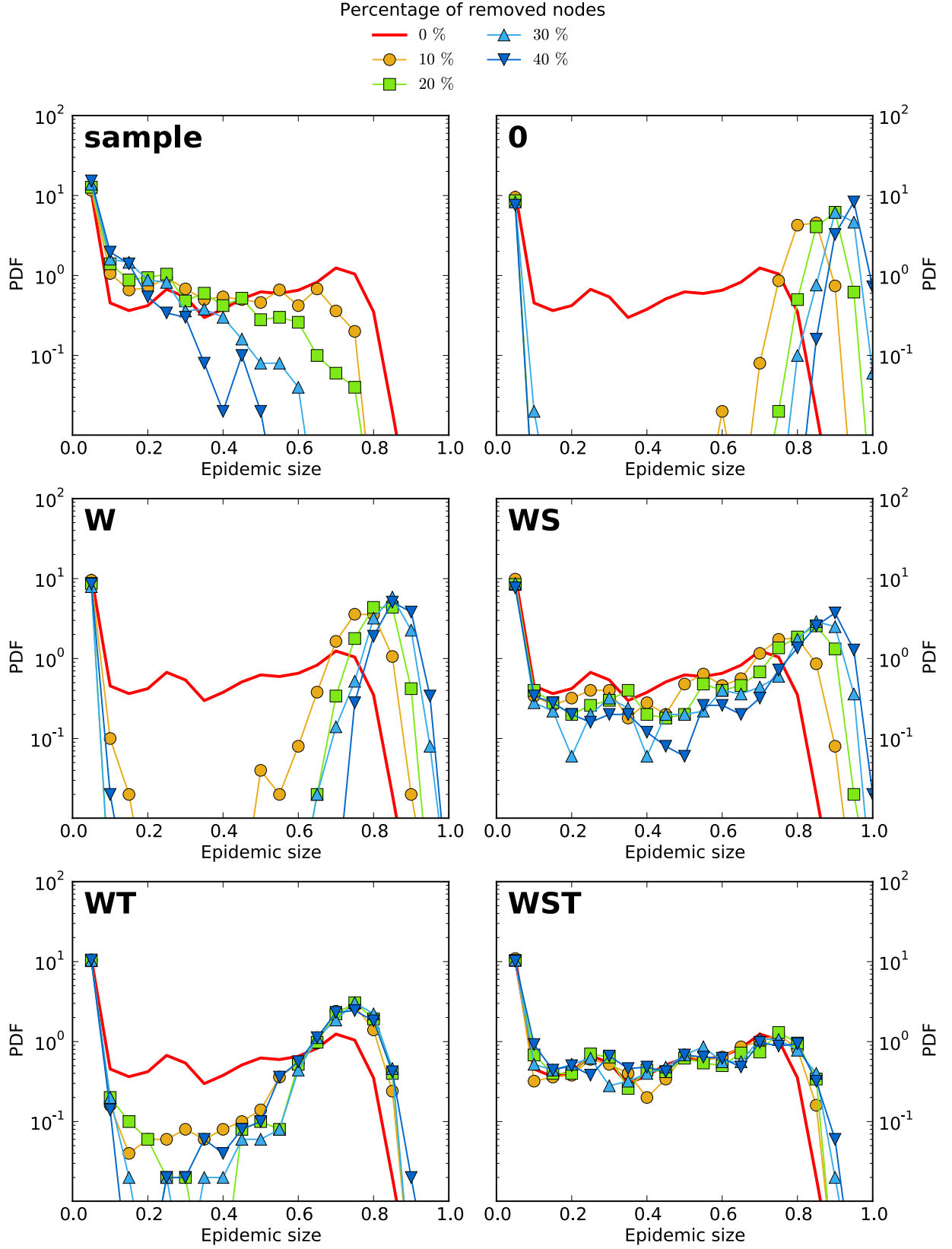


FIG. 3. **SIR simulations for the Thiers13 (high school) case.** We compare the outcome of SIR epidemic simulations performed on resampled (top left) and reconstructed contact networks, for different methods of reconstruction. We plot the distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of resampled (sample) and reconstructed contact networks, for different values of the fraction  $f$  of nodes removed, and for the 5 reconstruction methods described in the text (0, W, WS, WT, WST). The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 100$ . The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.

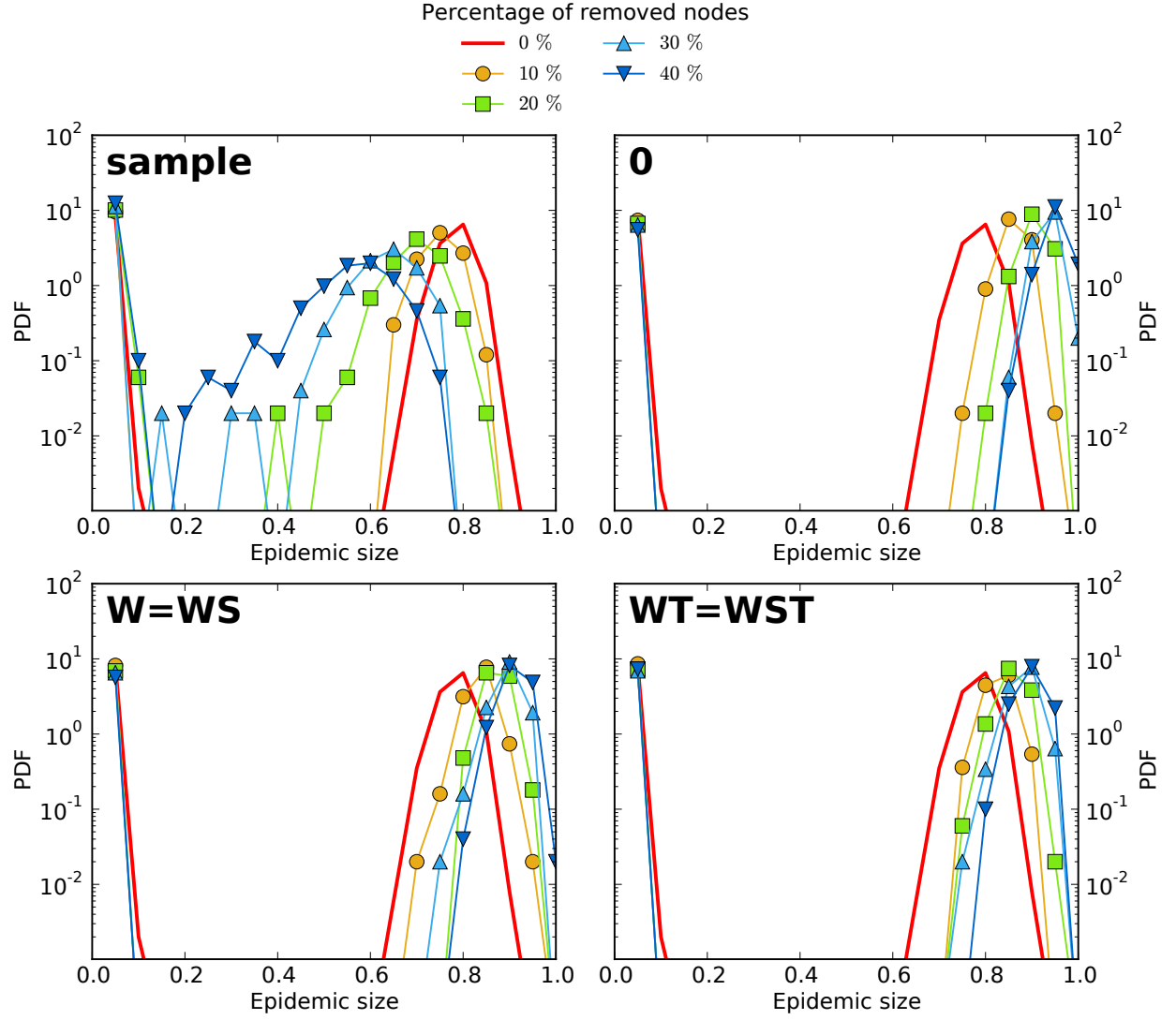


FIG. 4. **SIR simulations for the SFHH (conference) case.** We compare the outcome of SIR epidemic simulations performed on resampled and reconstructed contact networks, for different methods of reconstruction. We plot the distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of resampled (sample) and reconstructed contact networks, for different values of the fraction  $f$  of nodes removed, and for three reconstruction methods described in the text (0, W, WT). In this case, as the population is not structured, methods W and WS on the one hand, WT and WST on the other hand, are equivalent. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 100$ . The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.

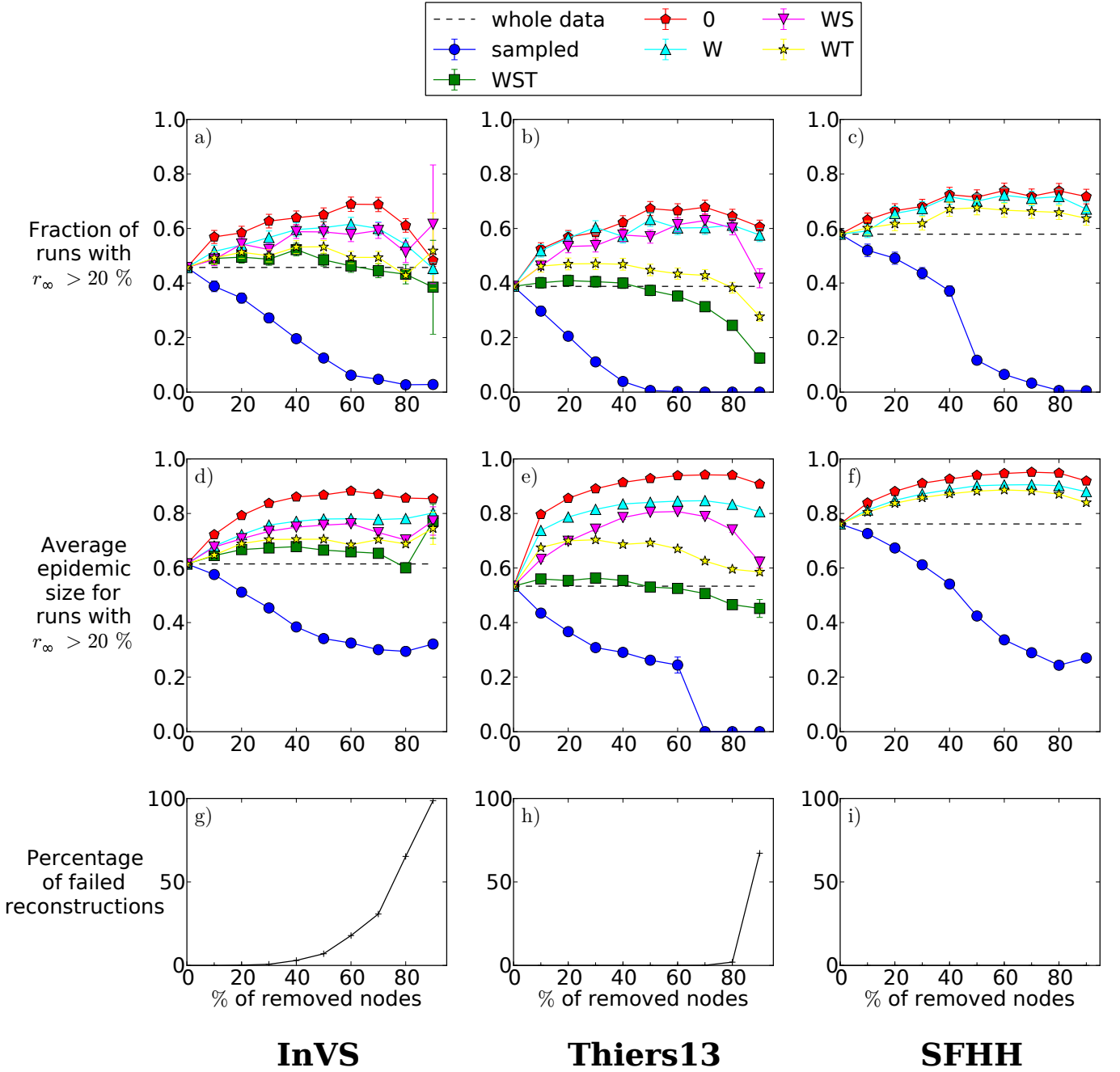


FIG. 5. **Accuracy of the different reconstruction methods.** We perform SIR epidemic simulations for each case, for different values of the fraction  $f$  of missing nodes, for both sampled networks and networks reconstructed with the different methods. We compare in each case, and as a function of  $f$ , the fraction of outbreaks that lead to a final fraction of recovered individuals  $r_\infty$  larger than 20% of the population (a, b, c), and the average size of these large outbreaks (d, e, f). The dashed lines give the corresponding values for simulations performed on the complete data sets. The different methods are: reconstruction conserving only the link density and the average weight of the resampled data (0); reconstruction conserving only the link density and the distribution of weights of the resampled data (W); reconstruction preserving, in addition to the W method, the group structure of the resampled data (WS); reconstruction conserving link density, distribution of weights and distributions of contact times, of inter-contact times and of numbers of contacts per link measured in the resampled data (WT); full method conserving all these properties (WST). We also plot as a function of  $f$  the failure rate of the WST algorithm, *i.e.*, the percentage of failed reconstructions (g, h, i). For the SFHH case, as the population is not structured into groups, methods W and WS are equivalent, as well as methods WT and WST; moreover, reconstruction is always possible. The SIR parameters are  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (InVS) or  $\beta/\mu = 100$  (Thiers13 and SFHH) and each point is averaged over 1,000 independent simulations.

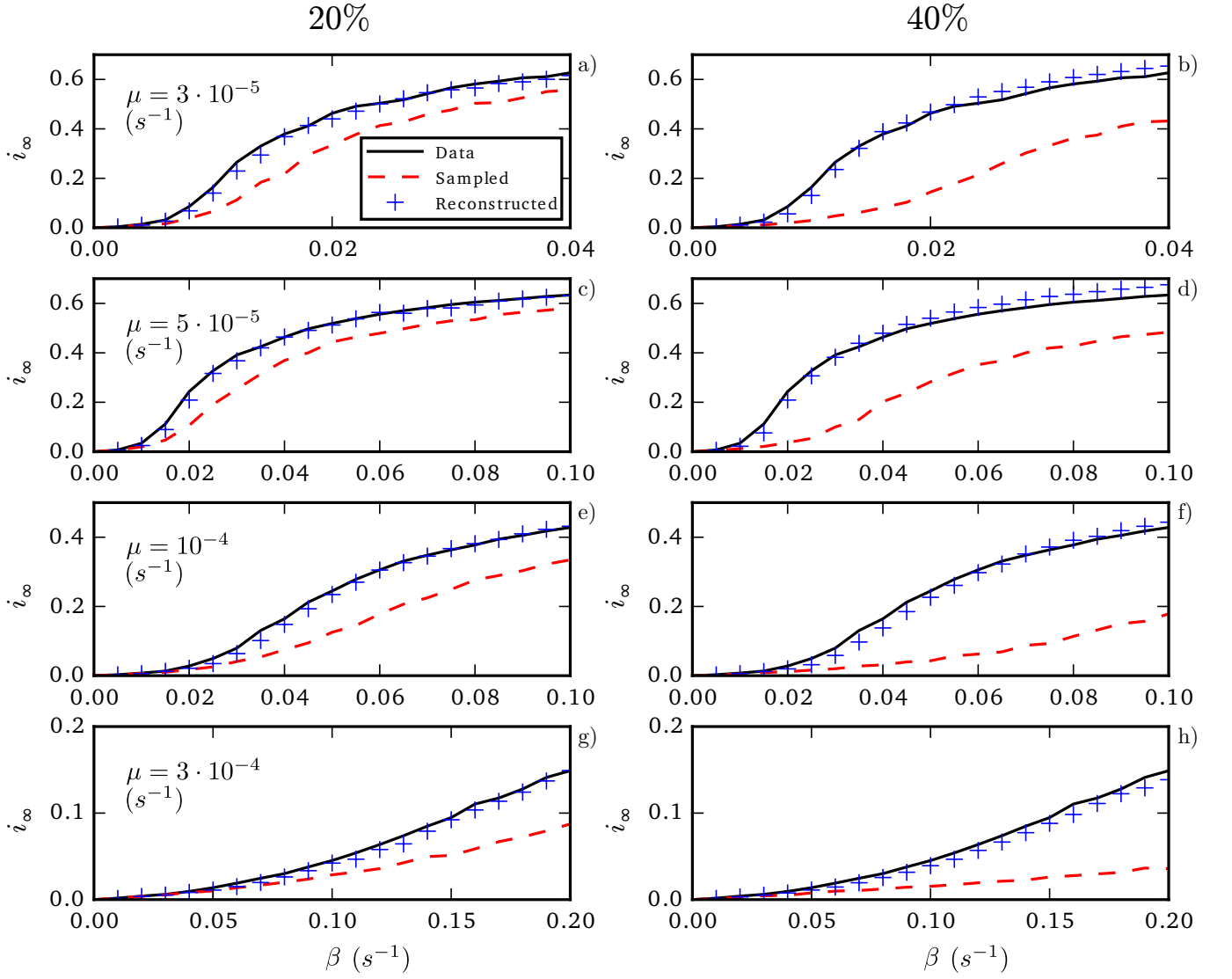


FIG. 6. **SIS simulations for the InVS (workplace) case.** We perform SIS epidemic simulations and report the phase diagram of the SIS model for the original, resampled and reconstructed contact networks. Each panel shows the stationary value  $i_\infty$  of the prevalence in the stationary state of the SIS model, computed as described in Methods, as a function of  $\beta$ , for several values of  $\mu$ . Simulations are performed in each case using either the complete data set (continuous lines), resampled data (dashed lines) or reconstructed contact networks using the WST method (pluses). The fraction of excluded nodes in the resampling is  $f = 20\%$  for a, c, e, g and  $f = 40\%$  for b, d, f, h.



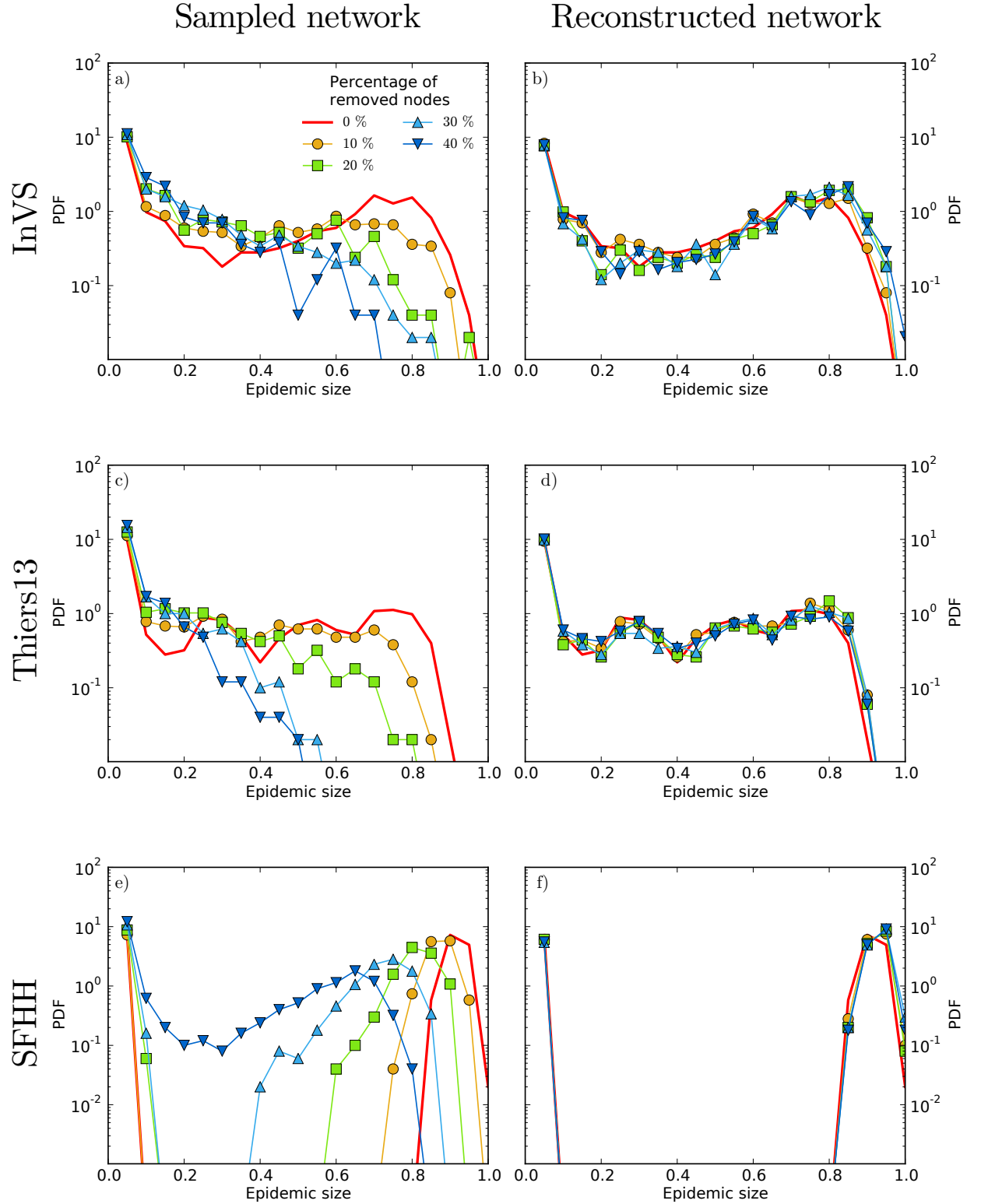


FIG. 7. **SIR simulations on shuffled data.** We compare of the outcome of SIR epidemic simulations performed on resampled and reconstructed contact networks, for shuffled data. We plot the distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (a, c, e) or reconstructed (b, d, f) contact networks, for different values of the fraction  $f$  of nodes removed. We use here the WST reconstruction method, and the data set considered consists in a CM-shuffled version (see Methods) of the original data, in which the shuffling procedure removes structural correlations of the contact network within each group. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (InVS) or  $\beta/\mu = 100$  (Thiers13 and SFHH). The case  $f = 0$  corresponds to simulations using the whole data set, *i.e.*, the reference (reshuffled data) case. For each value of  $f$ , 1,000 independent simulations were performed.

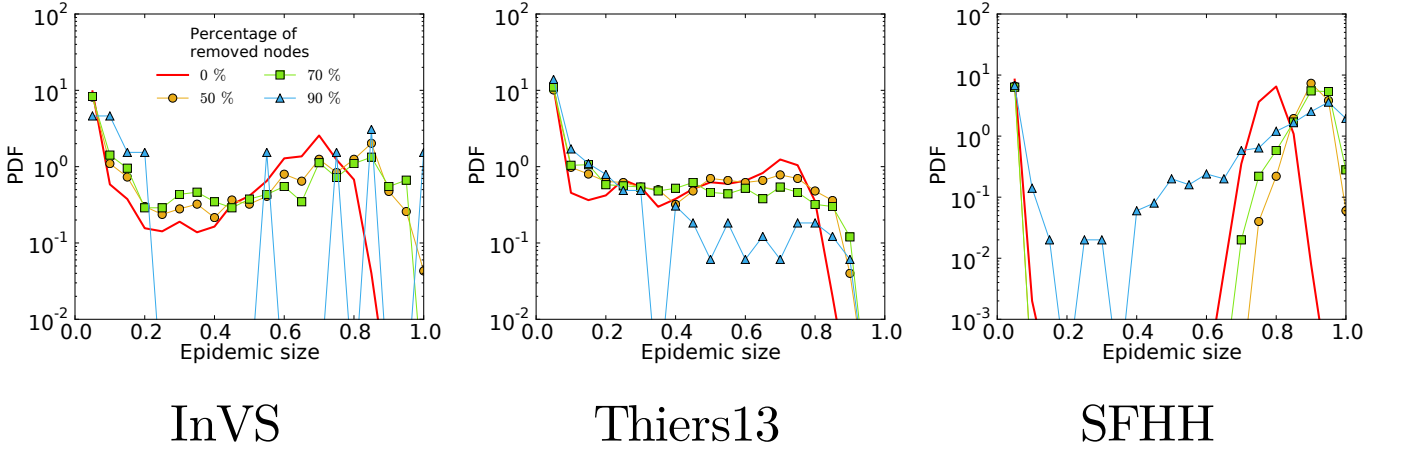
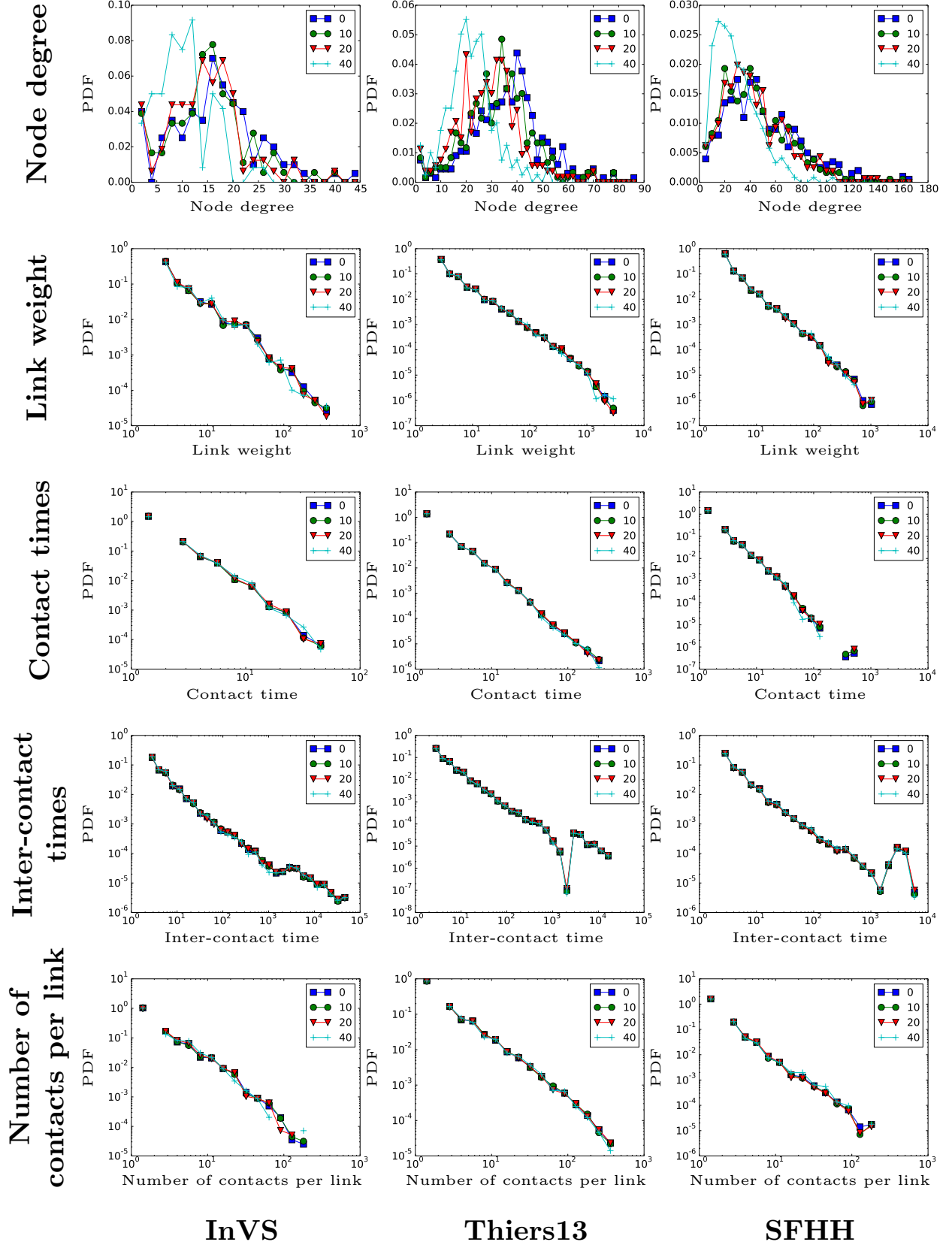
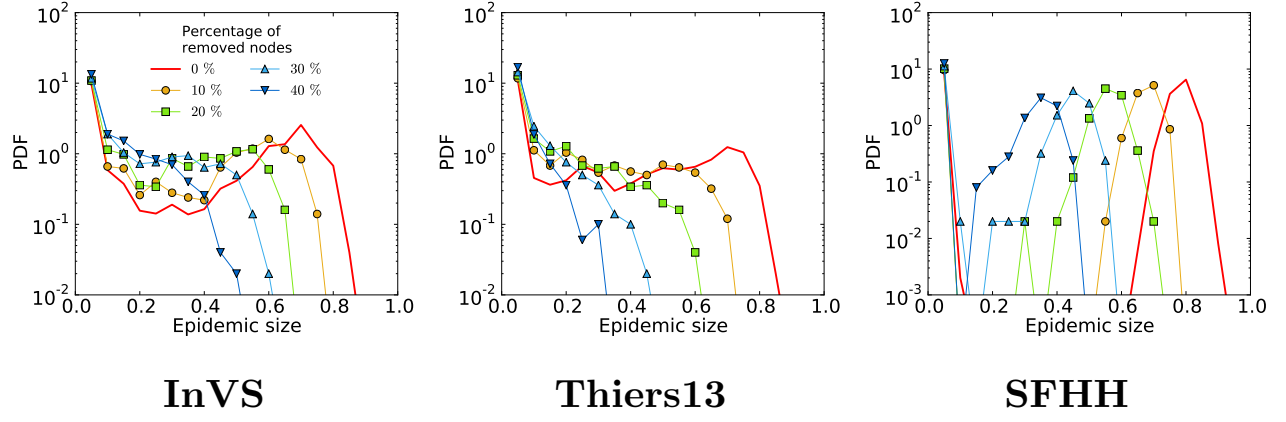


FIG. 8. **SIR simulations for very large numbers of missing nodes.** We simulate SIR processes on reconstructed contact networks for large values of the fraction  $f$  of removed nodes. We plot the distributions of epidemic sizes for simulations on reconstructed networks and on the whole data set (case  $f = 0$ ), for large values of the fraction  $f$  of removed nodes. Here  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (InVS) or  $\beta/\mu = 100$  (Thiers13 and SFHH) and 1,000 simulations were performed for each value of  $f$ . The distributions of epidemic sizes for simulations performed on resampled data sets are not shown since at these high values of  $f$ , almost no epidemics occur.

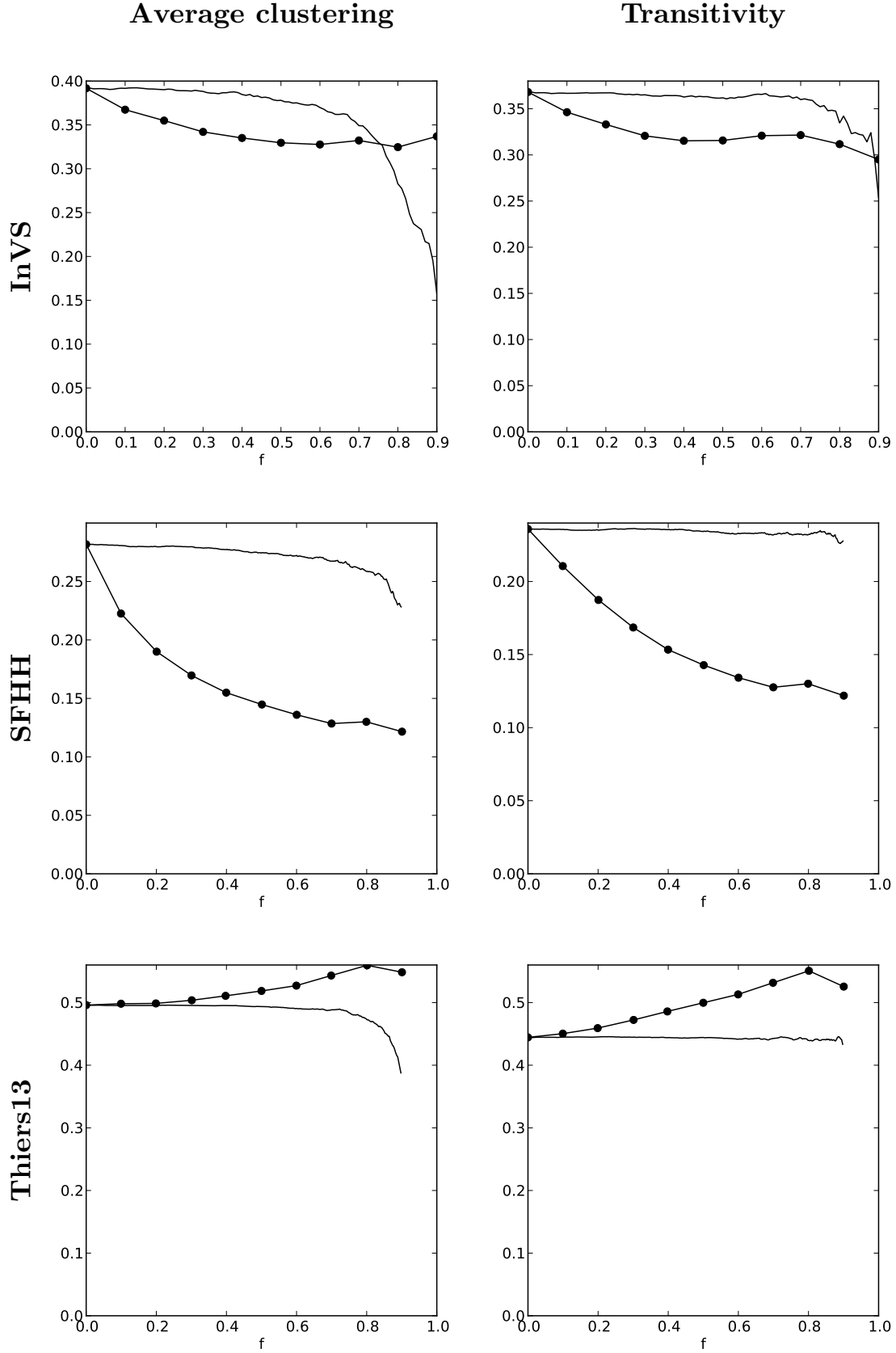
**SUPPLEMENTARY FIGURES**



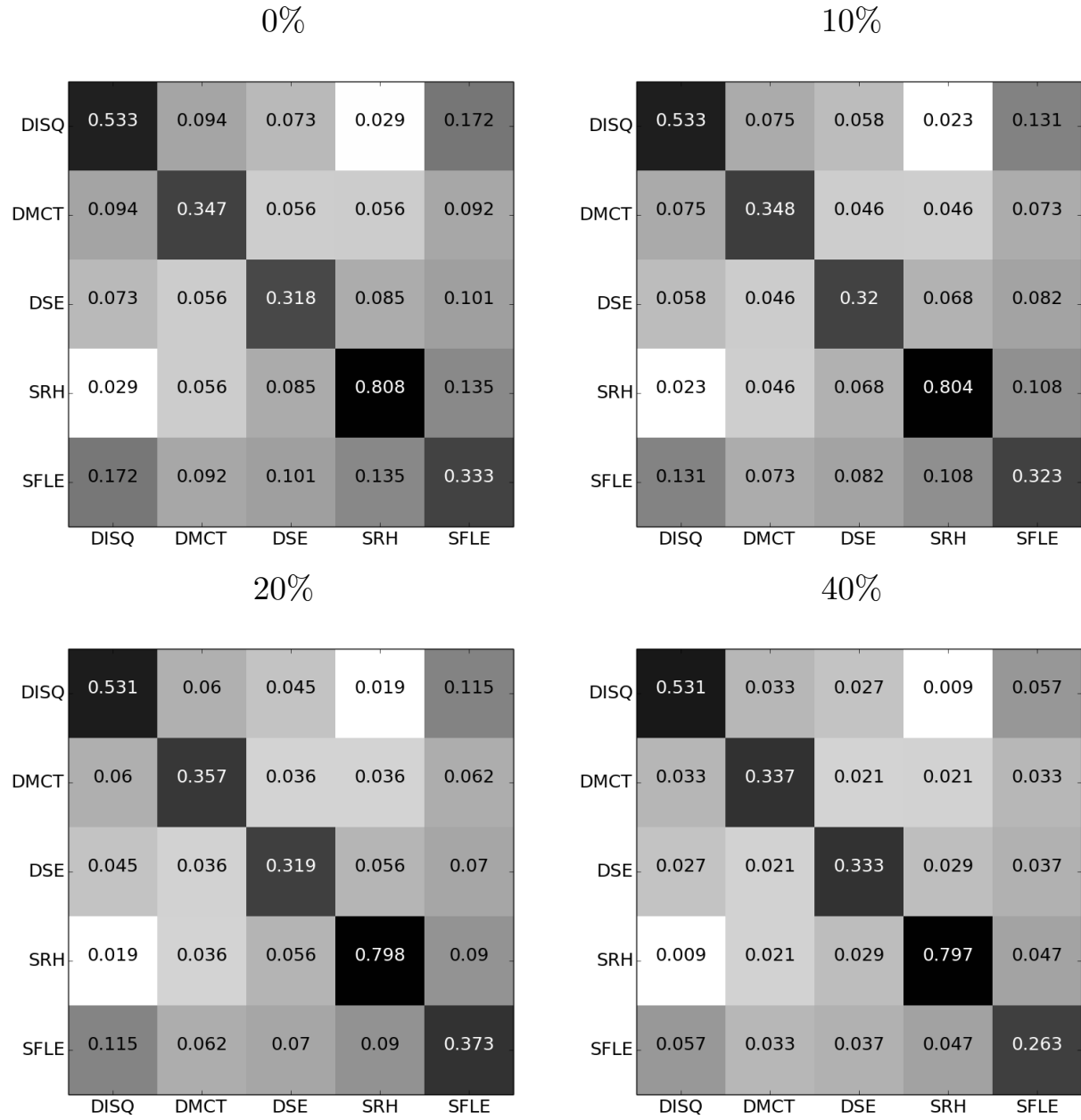
Supplementary Figure 9. **Effect of sampling on contact network properties.** Comparison of the distributions of structural (node degrees and link weights in the aggregated network of contacts) and temporal (contact durations, inter-contact times, number of contacts per link) properties of the contact networks, for different fractions  $f$  of removed nodes. For each value of  $f$ , the distributions are computed on a single realisation of the resampling.



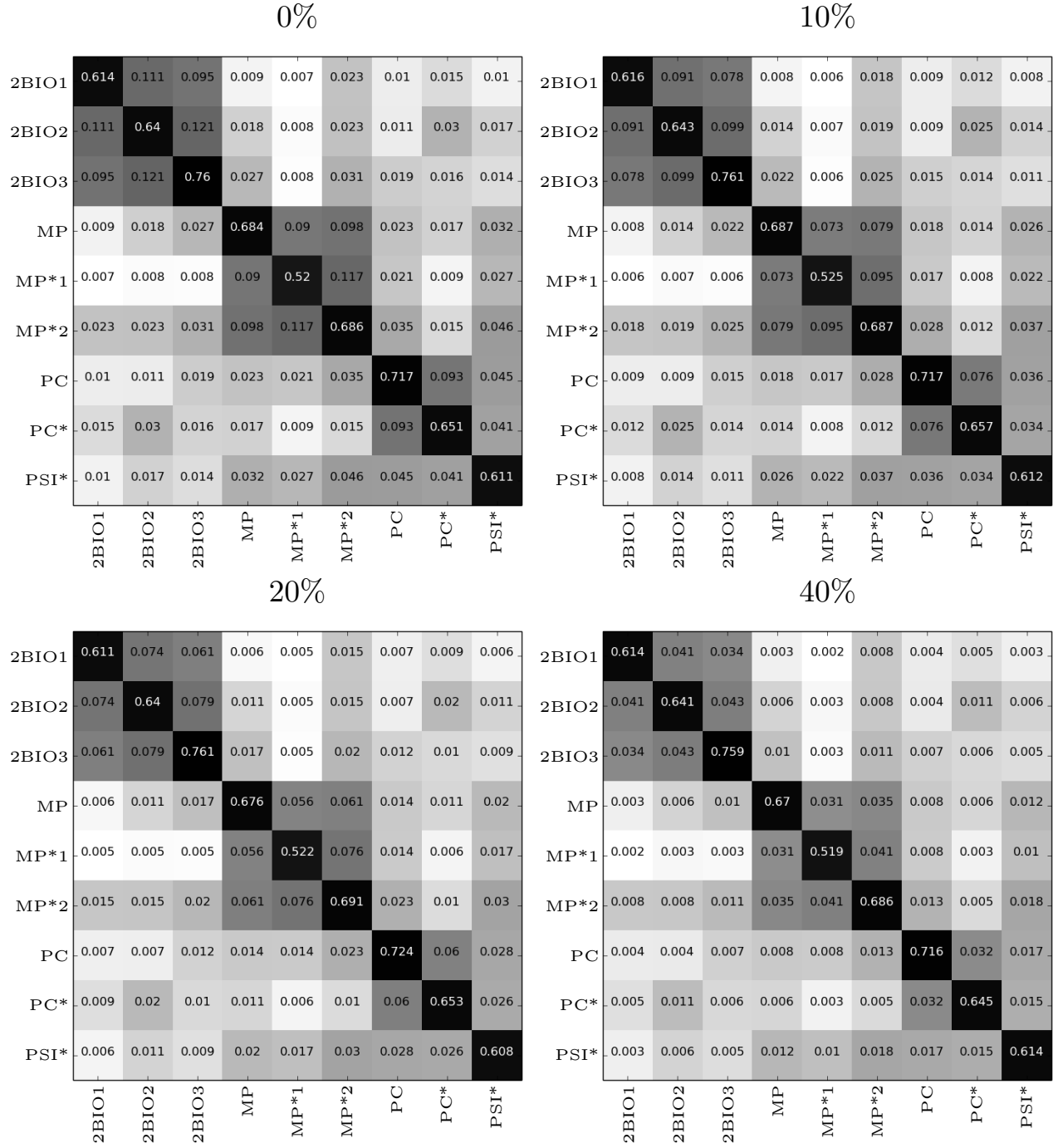
Supplementary Figure 10. **Effect of sampling on network density and on the similarity of contact matrices.** (Left) Density  $\rho$  of the aggregated network of contacts as a function of the fraction  $f$  of nodes excluded. The shaded areas represent mean  $\rho \pm$  s.e.m.. (Right) Median cosine similarities between the link density contact matrices (CML) of resampled and full data sets, as a function of  $f$ , for the structured populations (high school and workplace). Results are averaged, for each value of  $f$ , over 1,000 realisations for the density and over 100 realisations for the similarities.



Supplementary Figure 11. **Effect of sampling and reconstruction on the average clustering coefficient (left column) and on the network transitivity (right column).** The continuous lines show the evolution of clustering coefficient (left) and network transitivity when the fraction  $f$  of removed nodes increases. The full circles correspond to the same quantities for networks reconstructed using the **WST** method. Each point is averaged on 100 realisations.

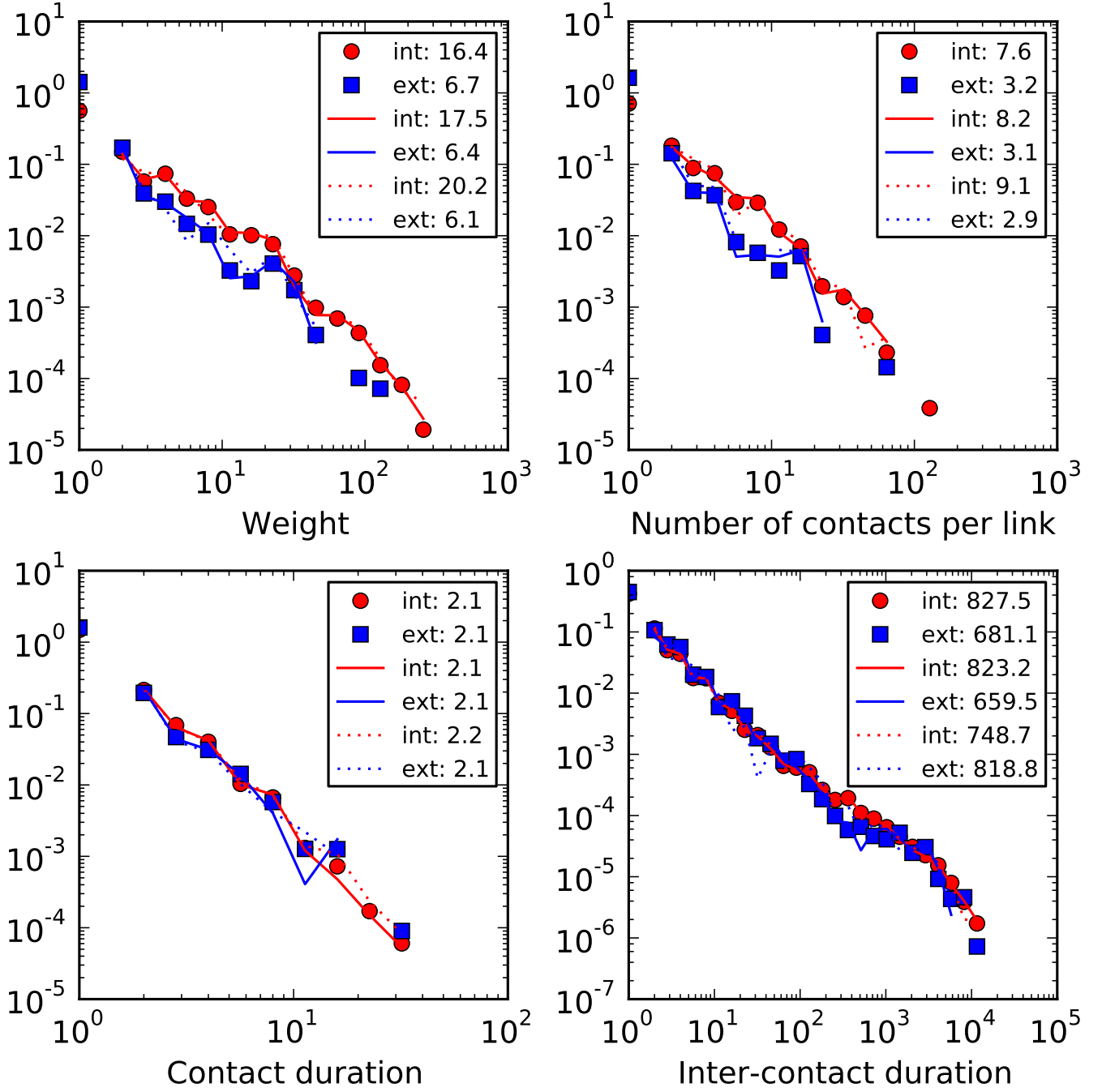


Supplementary Figure 12. **Effect of sampling: link density contact matrices (*InVS*)**. Comparison of link density contact matrices for the workplace, for different fractions of excluded nodes,  $f$ , with the original one ( $f = 0$ ). Each matrix element  $AB$  gives the number of links between nodes of department  $A$  and nodes of department  $B$  in the contact network, normalised by the maximum possible number of such links. For each value of  $f$ , each matrix element is an average over 100 realisations of the sampling.

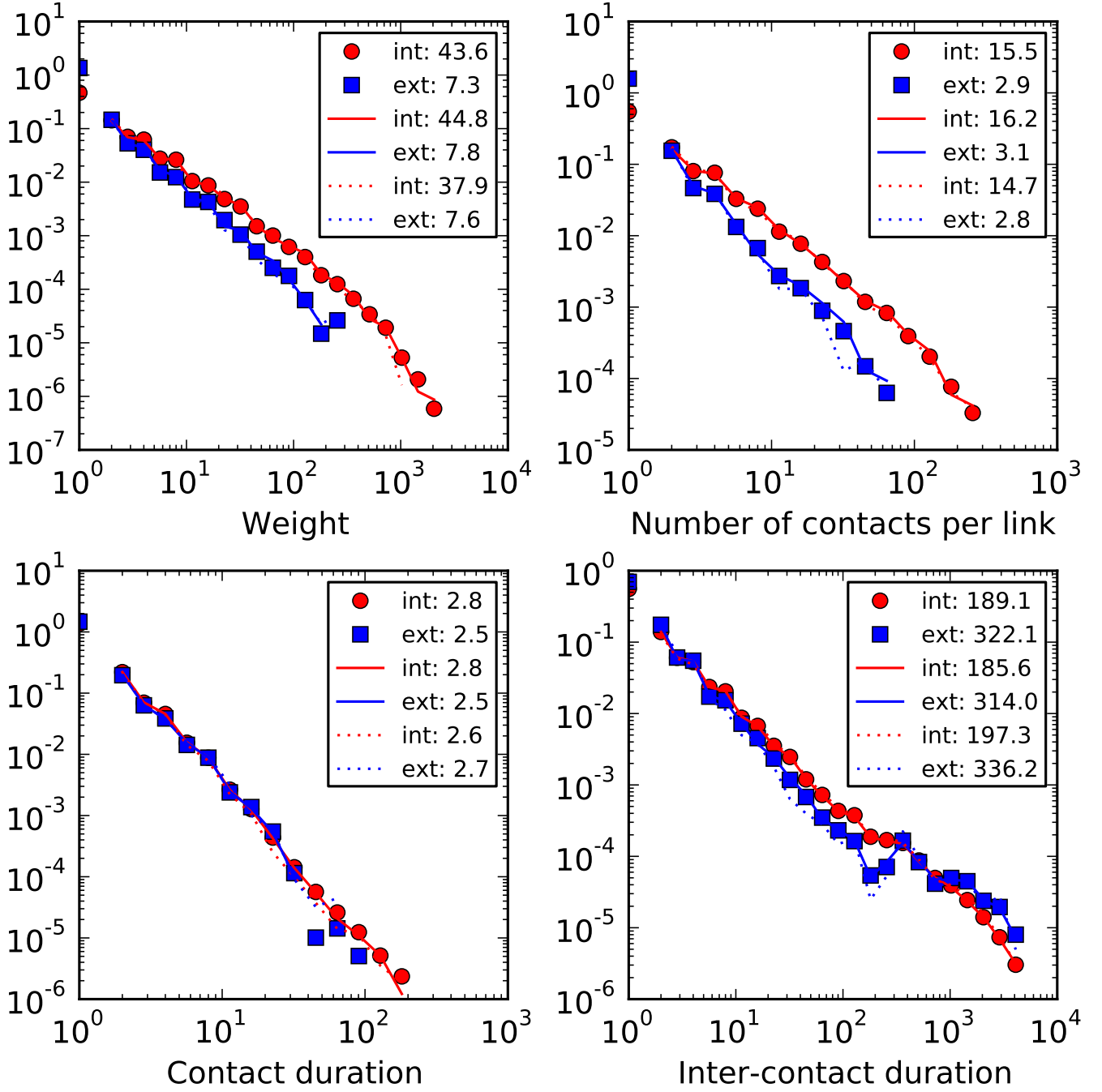


Supplementary Figure 13. **Effect of sampling: link density contact matrices (*Thiers13*)**. Comparison of the link density contact matrices for the high school, for different fractions  $f$  of excluded nodes, with the original one ( $f = 0$ ). Each matrix element  $AB$  gives the number of links between nodes of class  $A$  and nodes of class  $B$  in the contact network, normalised by the maximum possible number of such links. For each value of  $f$ , each matrix element is an average over 100 realisations of the sampling.

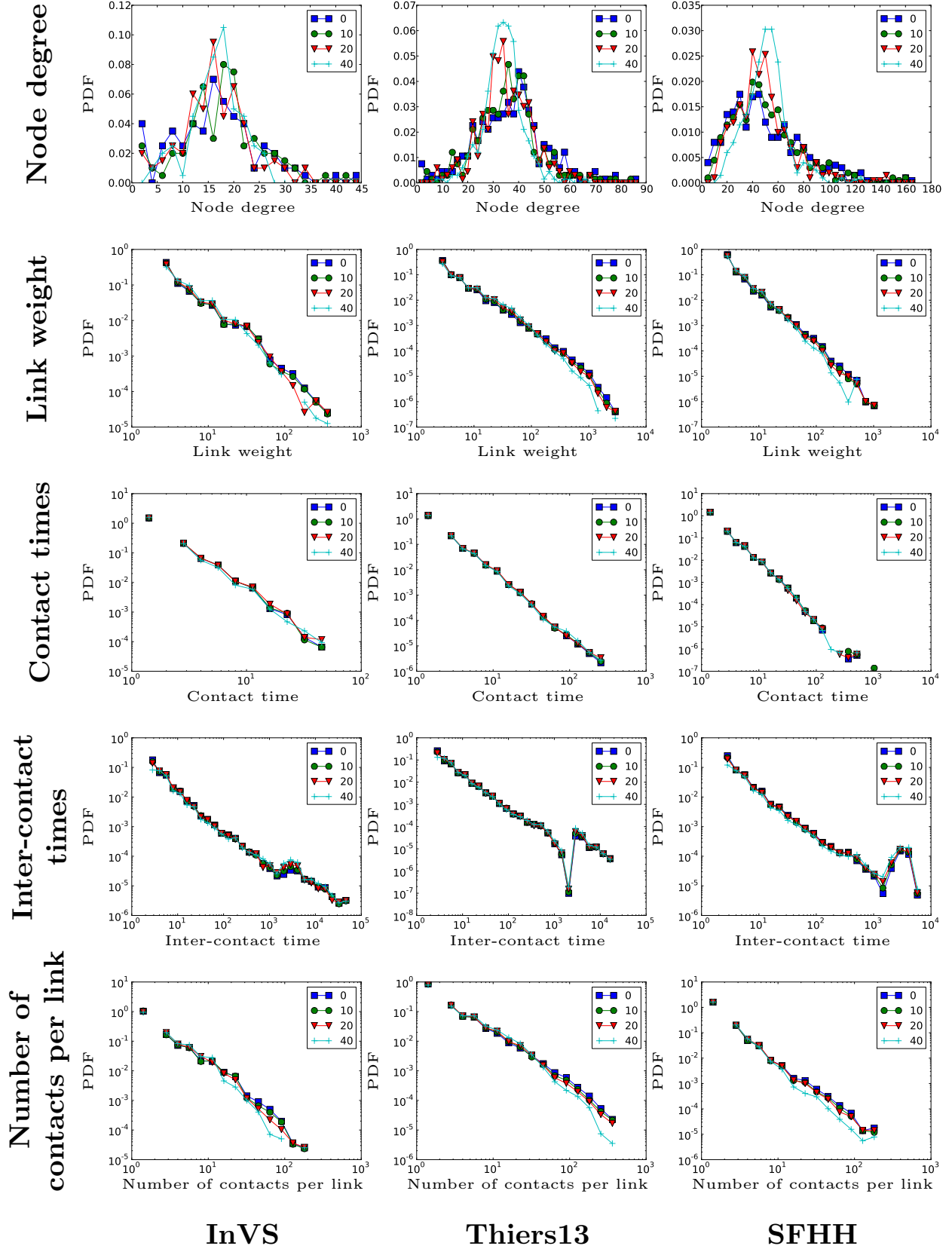




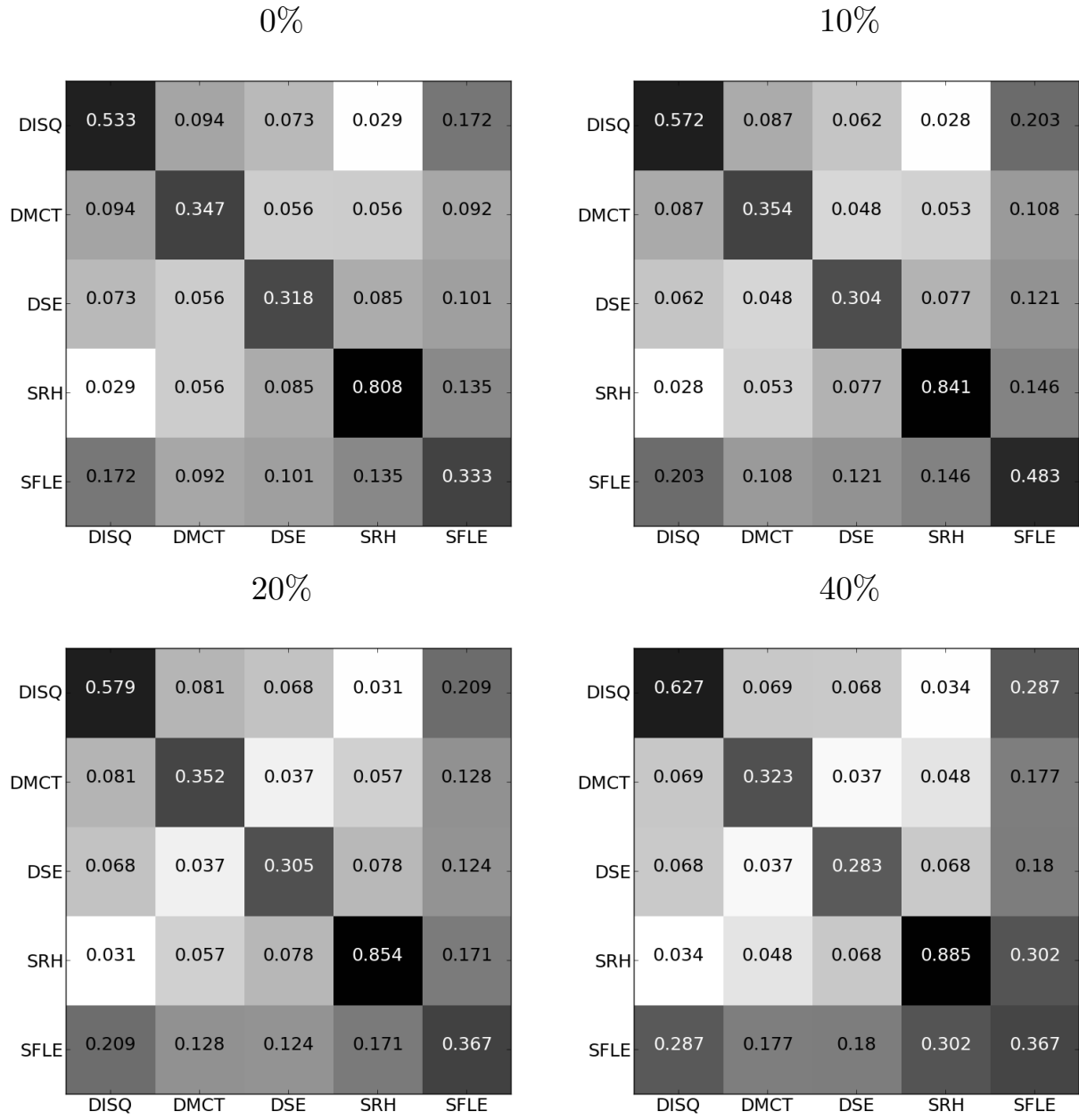
Supplementary Figure 14. **Distributions of temporal characteristics for internal (within groups) and external (between groups) contacts and links (*InVS* data).** Symbols are for the original data, full lines for resampled data with  $f = 20\%$ , dotted lines for  $f = 40\%$ . Legends give the average values for each distribution.



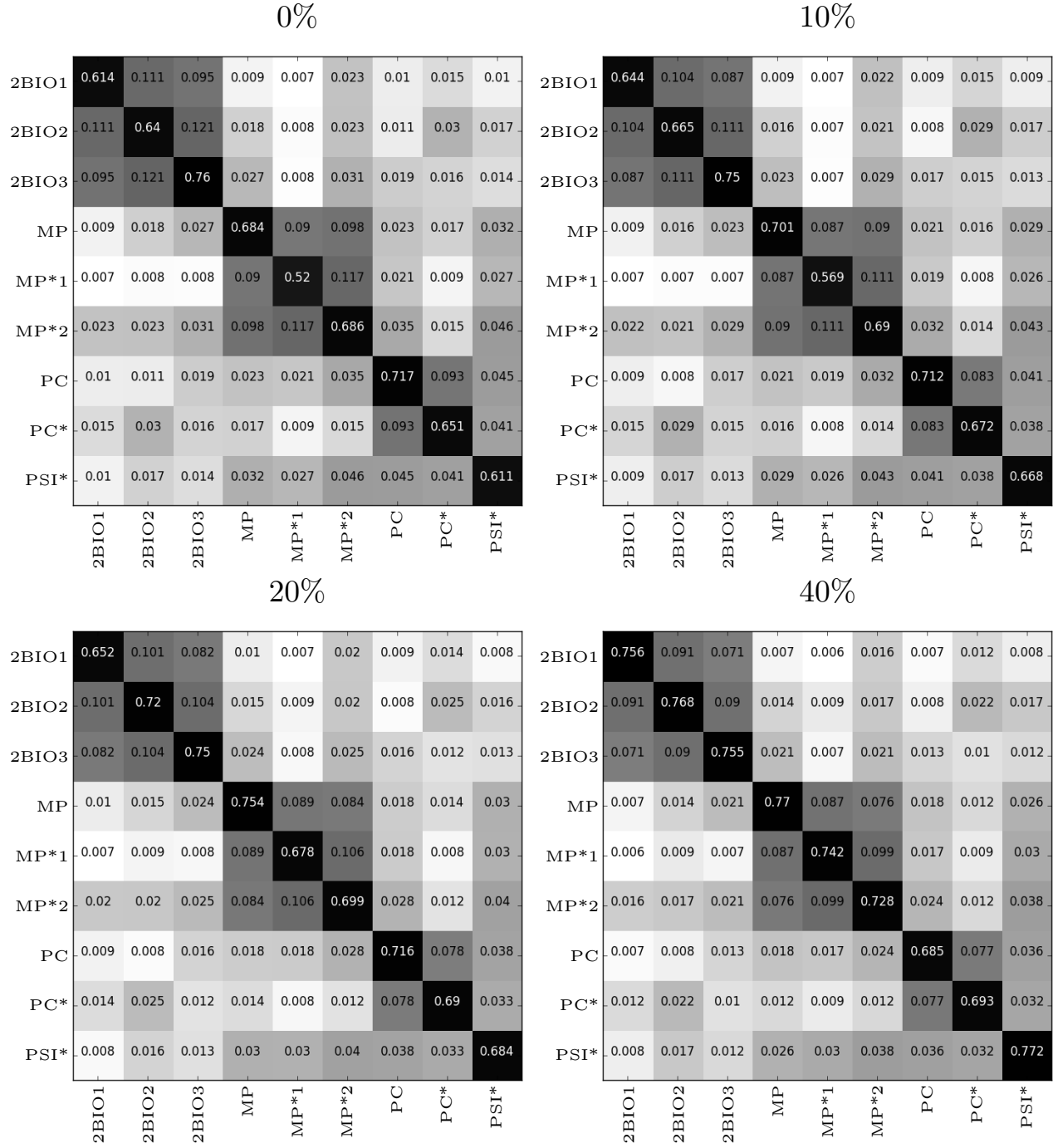
Supplementary Figure 15. **Distributions of temporal characteristics for internal (within groups) and external (between groups) contacts and links (*Thiers13* data).** Symbols are for the original data, full lines for resampled data with  $f = 20\%$ , dotted lines for  $f = 40\%$ . Legends give the average values for each distribution.



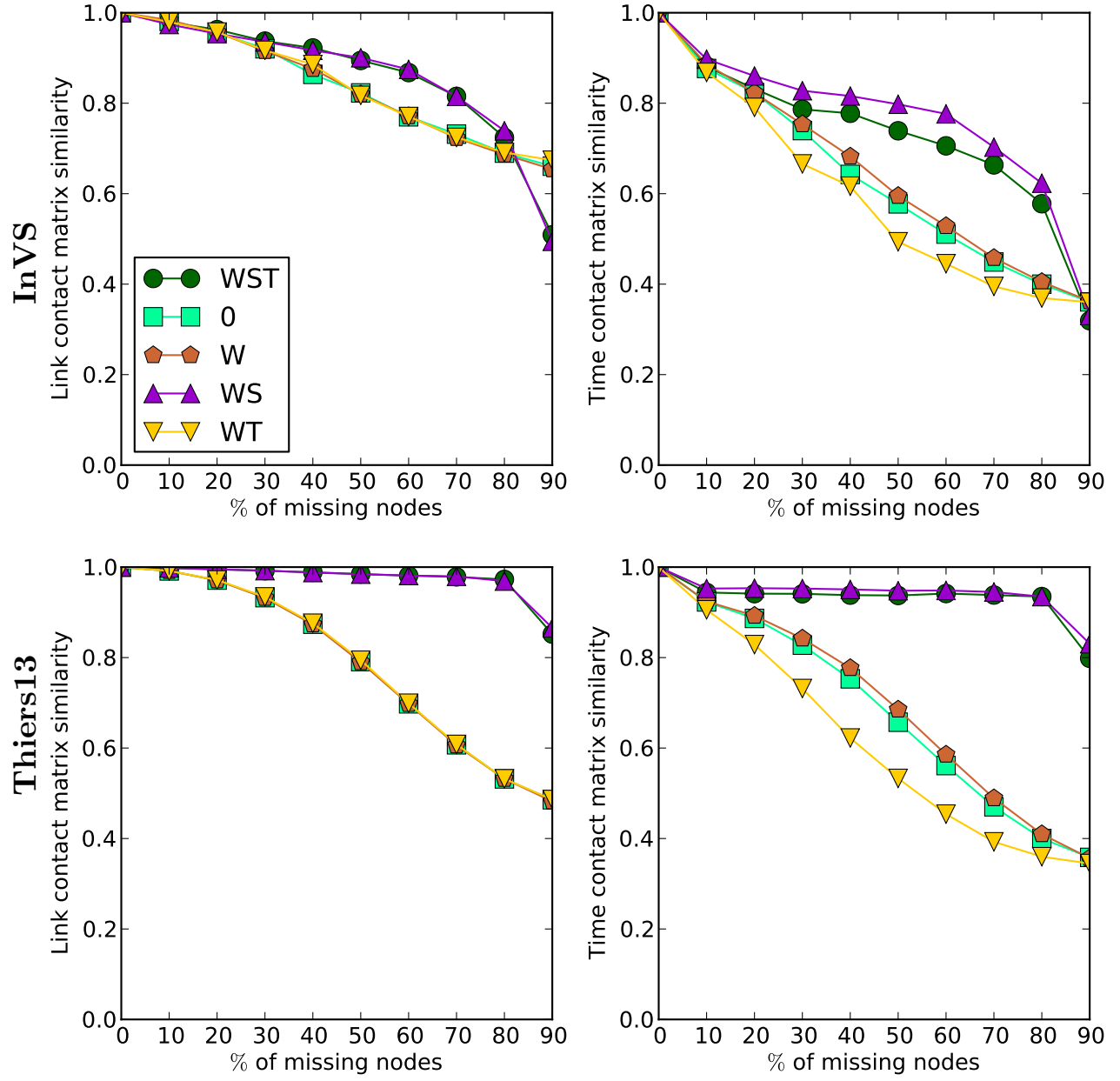
Supplementary Figure 16. **Properties of the reconstructed contact network.** Same as Fig. 9 but for the reconstructed networks: Distributions of structural (degrees and weights in the aggregated contact network) and temporal (contact times, inter-contact times, number of contacts per link) properties of the surrogate contact networks, for different fractions  $f$  of nodes excluded. For each value of  $f$ , the distributions are computed on a single reconstructed network.



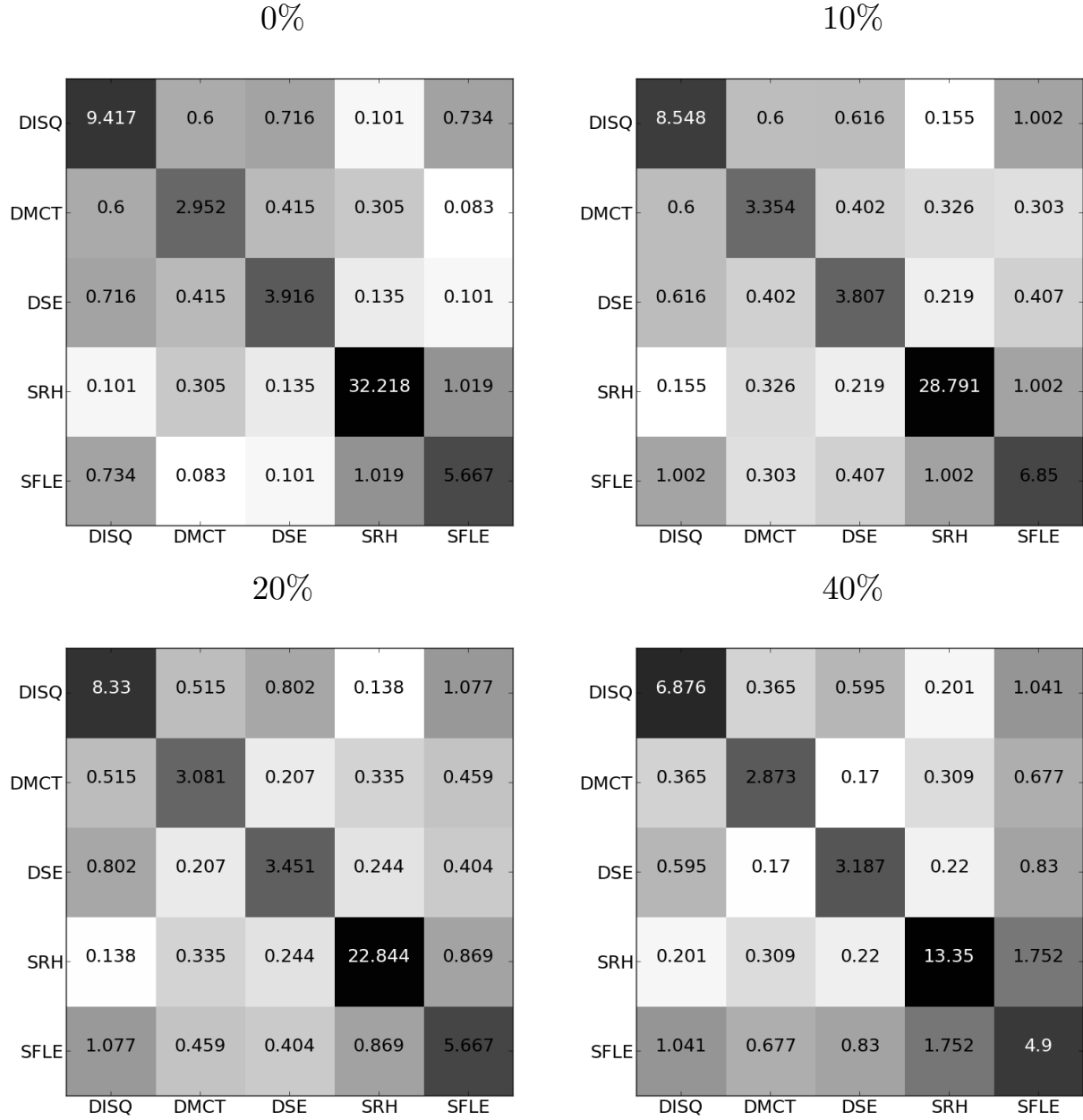
Supplementary Figure 17. **Properties of the reconstructed contact network: link density contact matrices (*InVS*).** Comparison of link density contact matrices for the reconstructed network of the workplace data, for different values of the fraction  $f$  of excluded nodes, with the original one ( $f = 0$ ). For each value of  $f$ , each matrix element is an average over 100 realisations of the sampling.



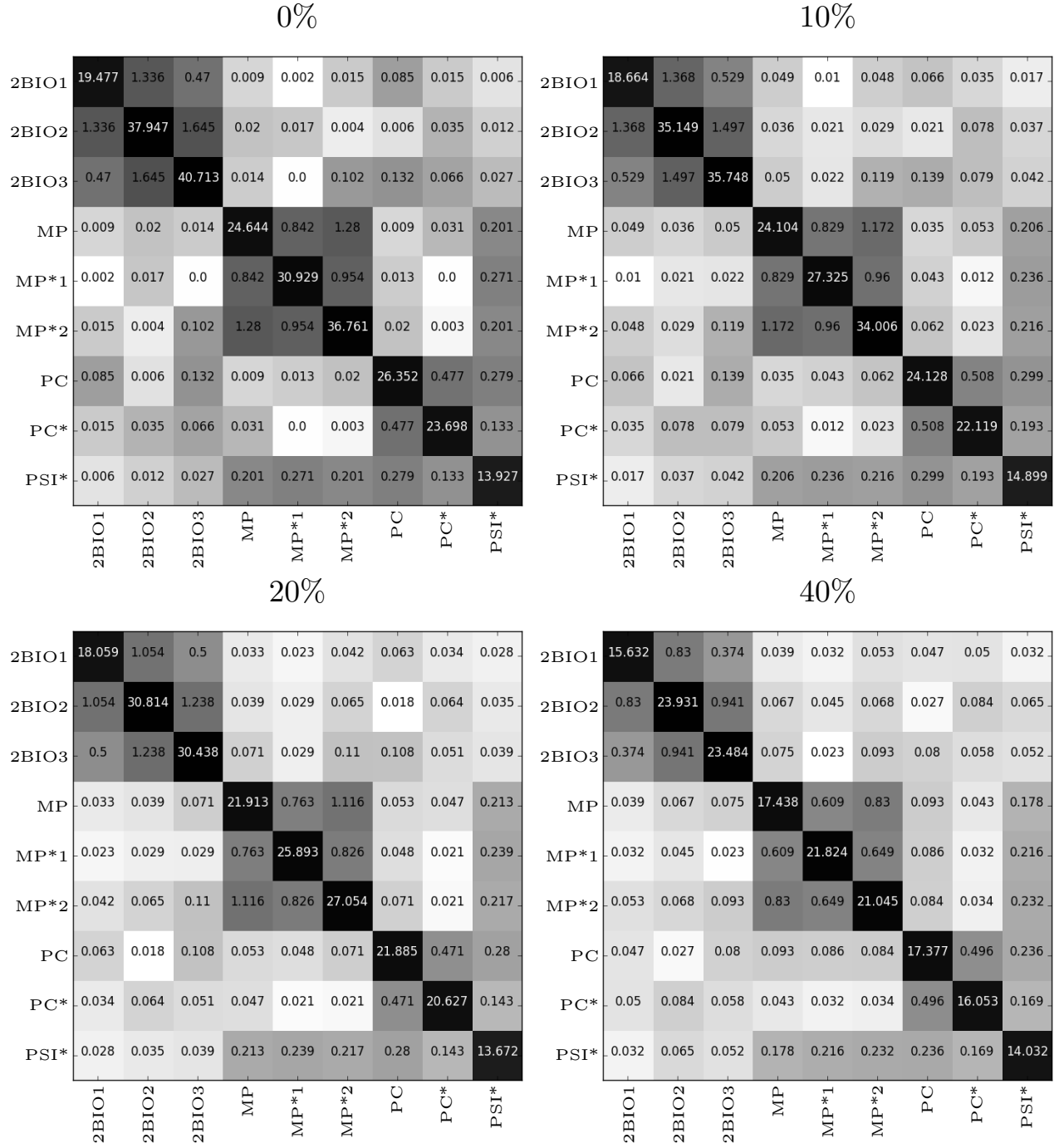
Supplementary Figure 18. **Properties of the reconstructed contact network: link density contact matrices (*Thiers13*)**. Comparison of link density contact matrices for the reconstructed network of the high school data, for different values of the fraction  $f$  of excluded nodes, with the original one ( $f = 0$ ). For each value of  $f$ , each matrix element is an average over 100 realisations of the sampling.



Supplementary Figure 19. **Similarity of contact matrices for different reconstruction methods** Median cosine similarity between the link density and contact time density contact matrices computed between the reconstructed network and for the original contact matrices, as a function of the fraction  $f$  of removed nodes. For each value of  $f$ , the median is computed over 100 realisations of the reconstruction.

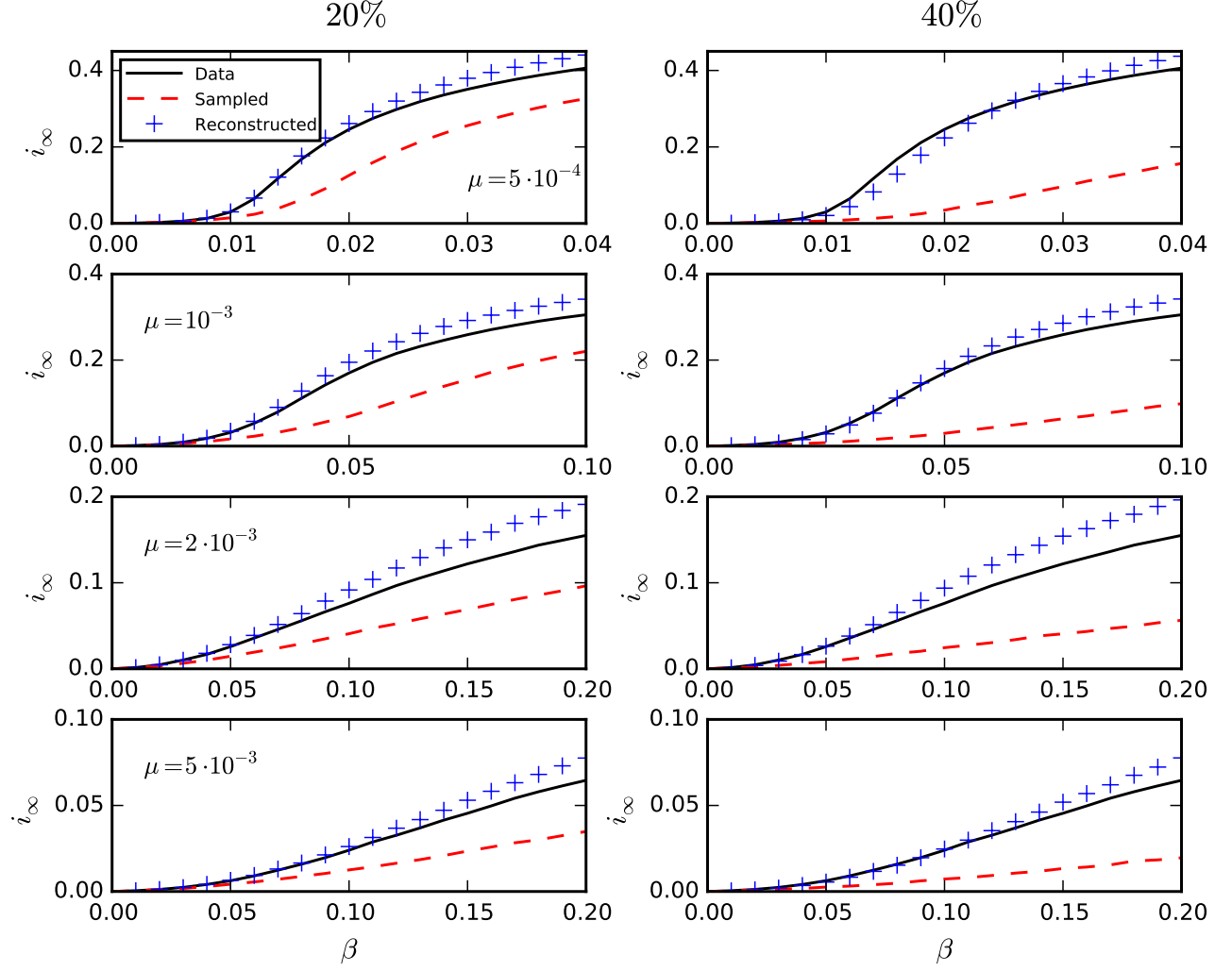


Supplementary Figure 20. **Properties of the reconstructed contact network: time density contact matrices (*InVS*).** Comparison of the contact time density contact matrices for the reconstructed network of the workplace data, for different fractions of excluded nodes,  $f$ , with the original one ( $f = 0$ ). Each matrix element  $AB$  gives the average time spent in contact between a node of department  $A$  and a node of department  $B$ . For each value of  $f$ , each matrix element is an average over 100 realisations of the sampling.

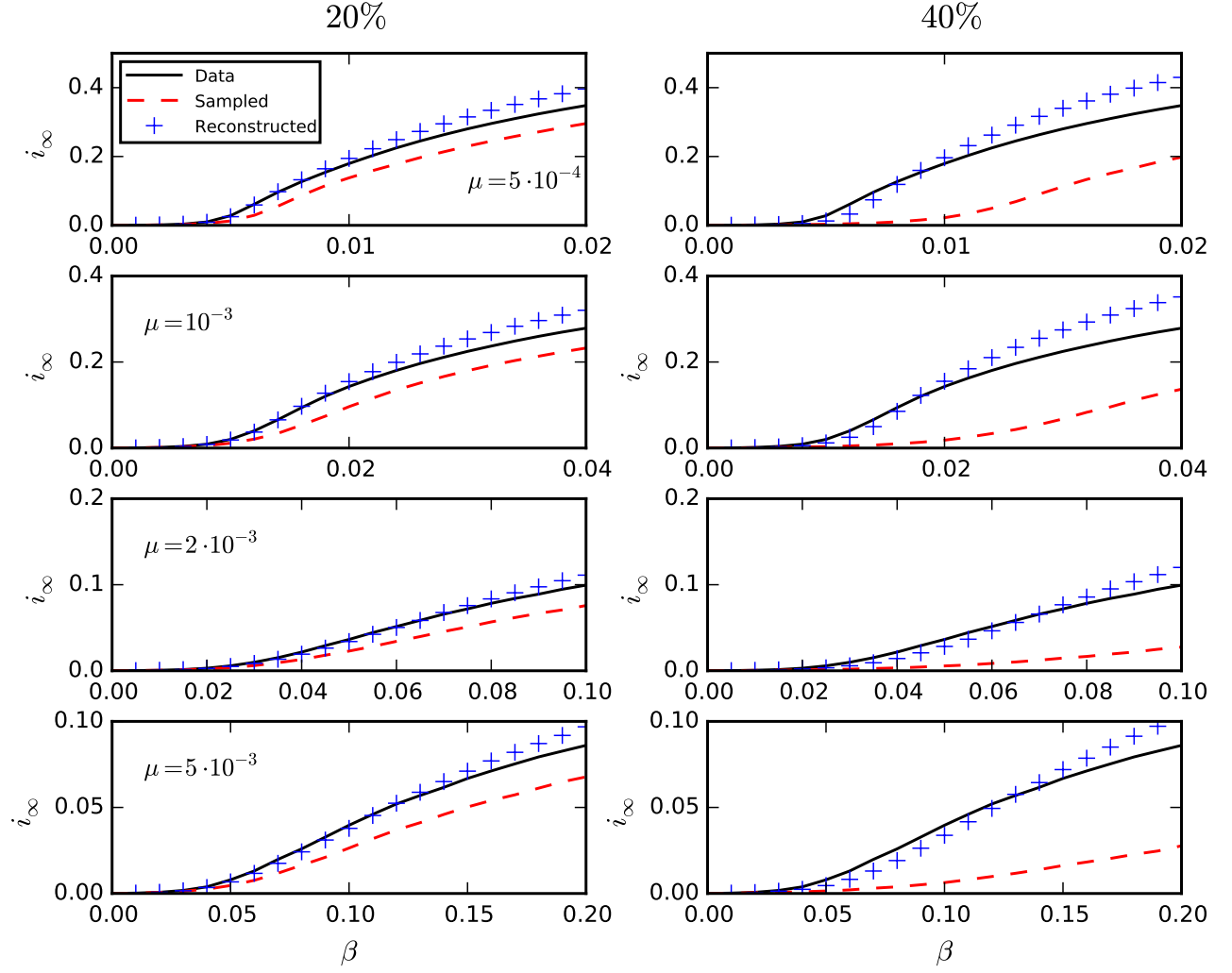


Supplementary Figure 21. **Properties of the reconstructed contact network: time density contact matrices (*Thiers13*)**. Contact time density contact matrices for the reconstructed network of the high school data, for different fractions of nodes excluded,  $f$ . Each matrix element  $AB$  gives the average time spent in contact between a node of class  $A$  and a node of class  $B$ . For each value of  $f$ , each matrix element is an average over 100 realisations of the sampling.

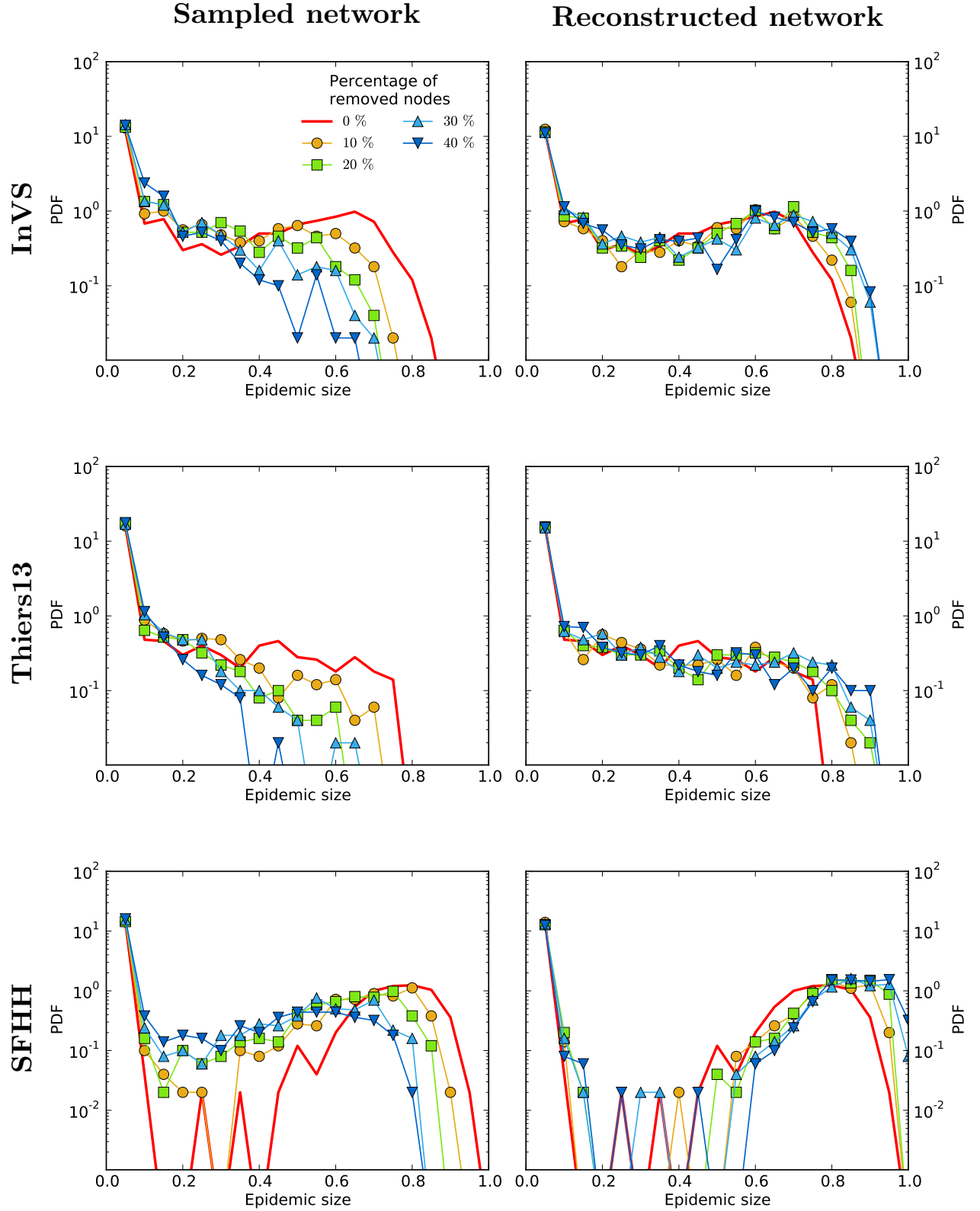




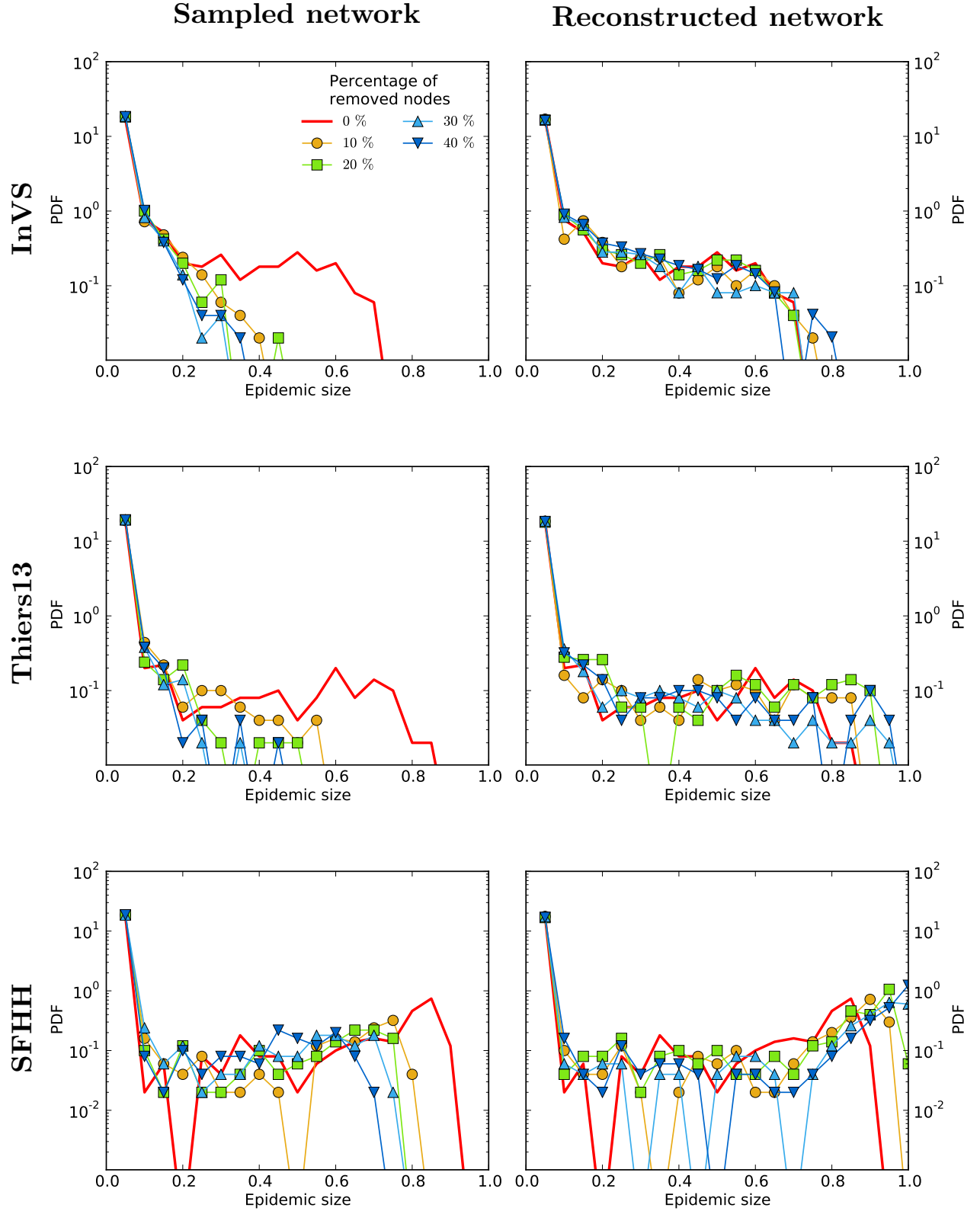
Supplementary Figure 22. **Phase diagram of the SIS model for original, resampled and reconstructed contact networks (*Thiers13* data set).** Each panel shows the stationary value  $i_\infty$  of the prevalence in the stationary state of the SIS model, computed as described in the Methods section, as a function of  $\beta$ , for several values of  $\mu$ . Here we consider the example of the *Thiers13* data set. The epidemic threshold corresponds to the transition between  $i_\infty = 0$  and  $i_\infty > 0$ . The prevalence curves are computed in each case using either the whole data set (continuous lines), resampled data (dashed lines) or reconstructed contact networks (pluses). The fraction of excluded nodes in the resampling is  $f = 20\%$  for the left column and  $f = 40\%$  for the right column.



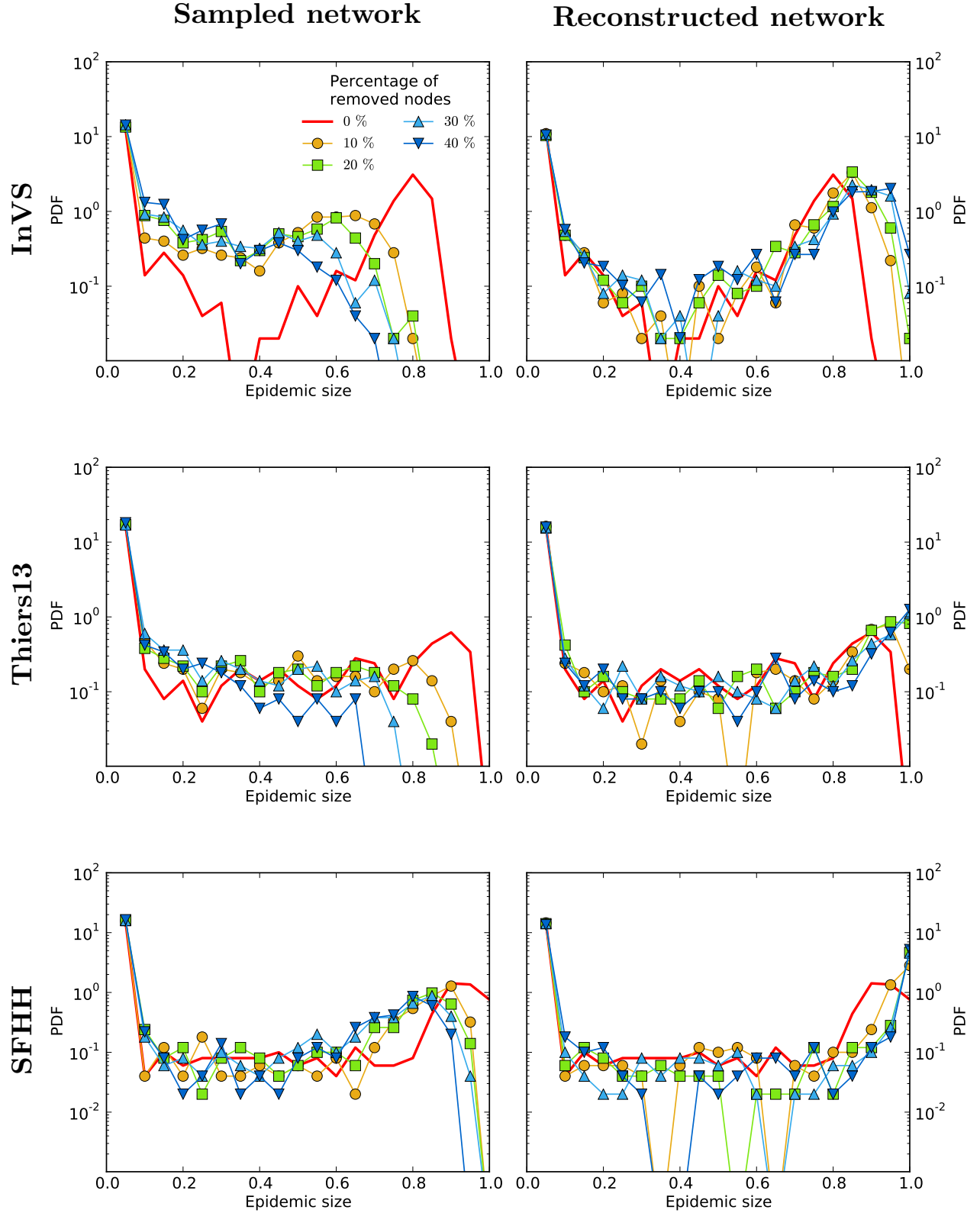
Supplementary Figure 23. **Phase diagram of the SIS model for original, resampled and reconstructed contact networks (*SFHH* data set).** Same as Fig. 22 for the *SFHH* (conference) data set.



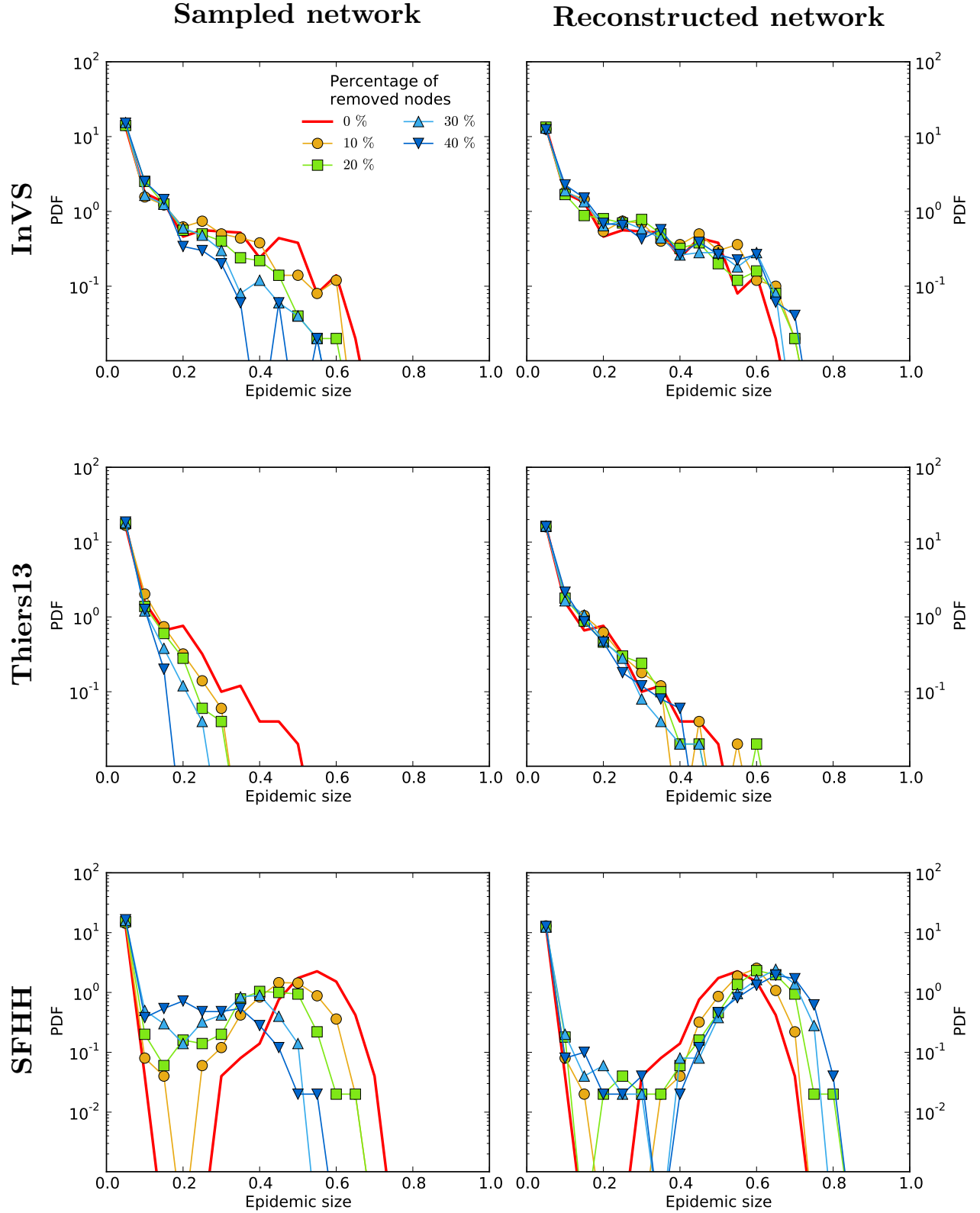
Supplementary Figure 24. **Outcome of SIR epidemic simulations on resampled and reconstructed networks for different parameter values.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, using the **WST** method, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.004$  and  $\beta/\mu = 1000$  (*InVS*) or  $\beta/\mu = 100$  (*Thiers13* and *SFHH*). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.



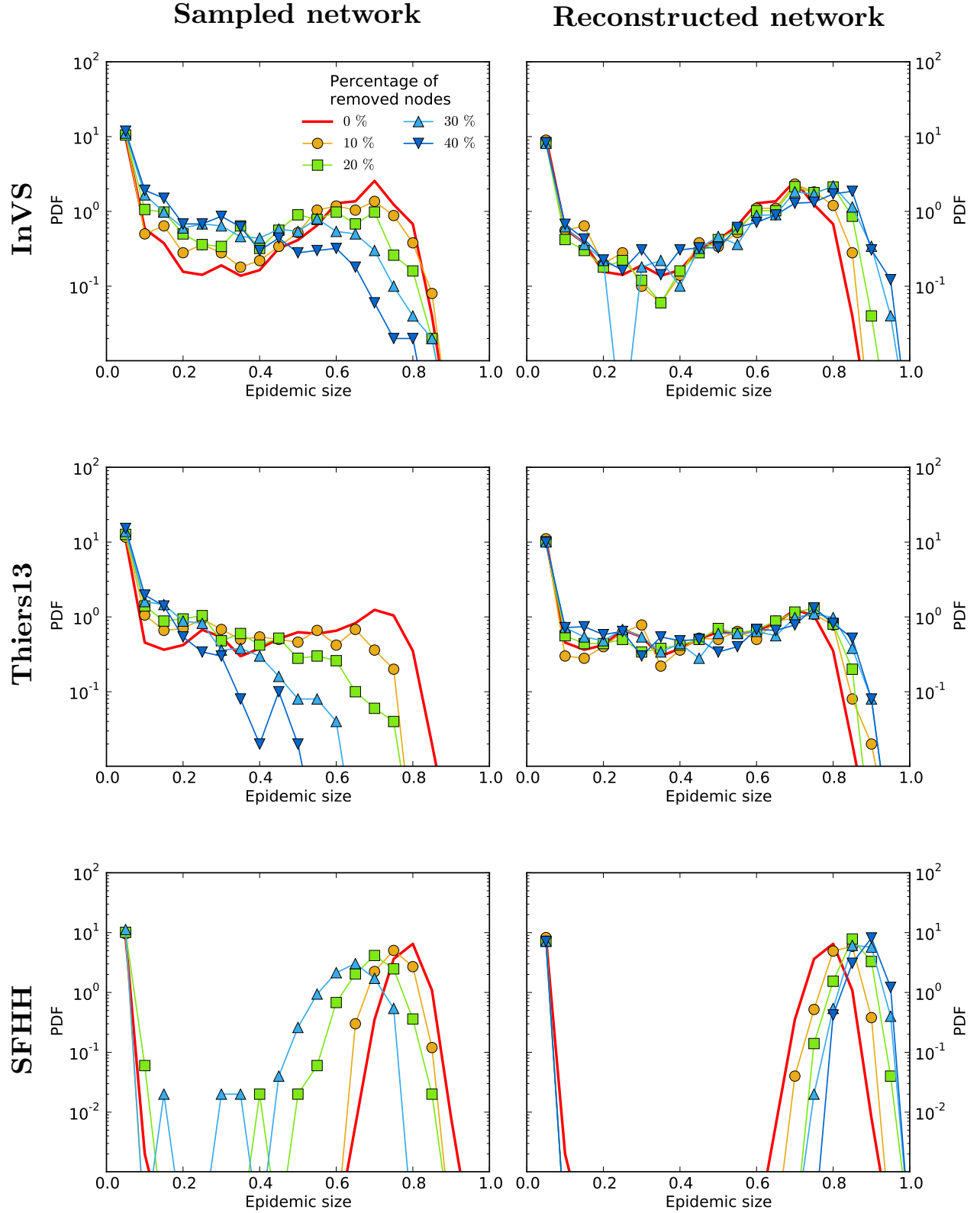
Supplementary Figure 25. **Outcome of SIR epidemic simulations on resampled and reconstructed networks for different parameter values.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, using the **WST** method, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.04$  and  $\beta/\mu = 1000$  (*InVS*) or  $\beta/\mu = 100$  (*Thiers13* and *SFHH*). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.



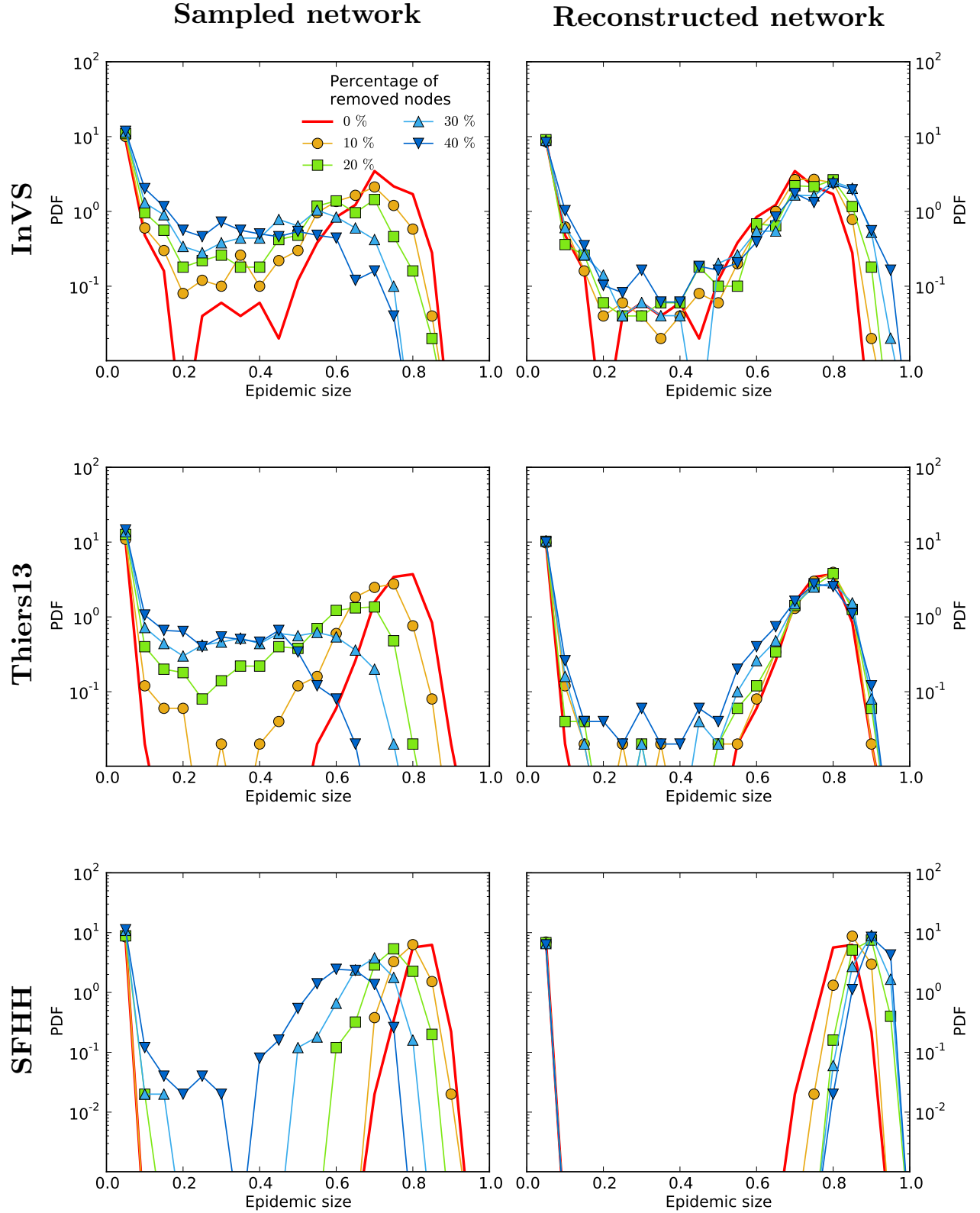
Supplementary Figure 26. **Outcome of SIR epidemic simulations on resampled and reconstructed networks for different parameter values.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, using the **WST** method, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.04$  and  $\beta/\mu = 4000$  (*InVS*) or  $\beta/\mu = 400$  (*Thiers13* and *SFHH*). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.



Supplementary Figure 27. **Outcome of SIR epidemic simulations on resampled and reconstructed networks for different parameter values.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, using the **WST** method, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 500$  (*InVS*) or  $\beta/\mu = 50$  (*Thiers13* and *SFHH*). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.

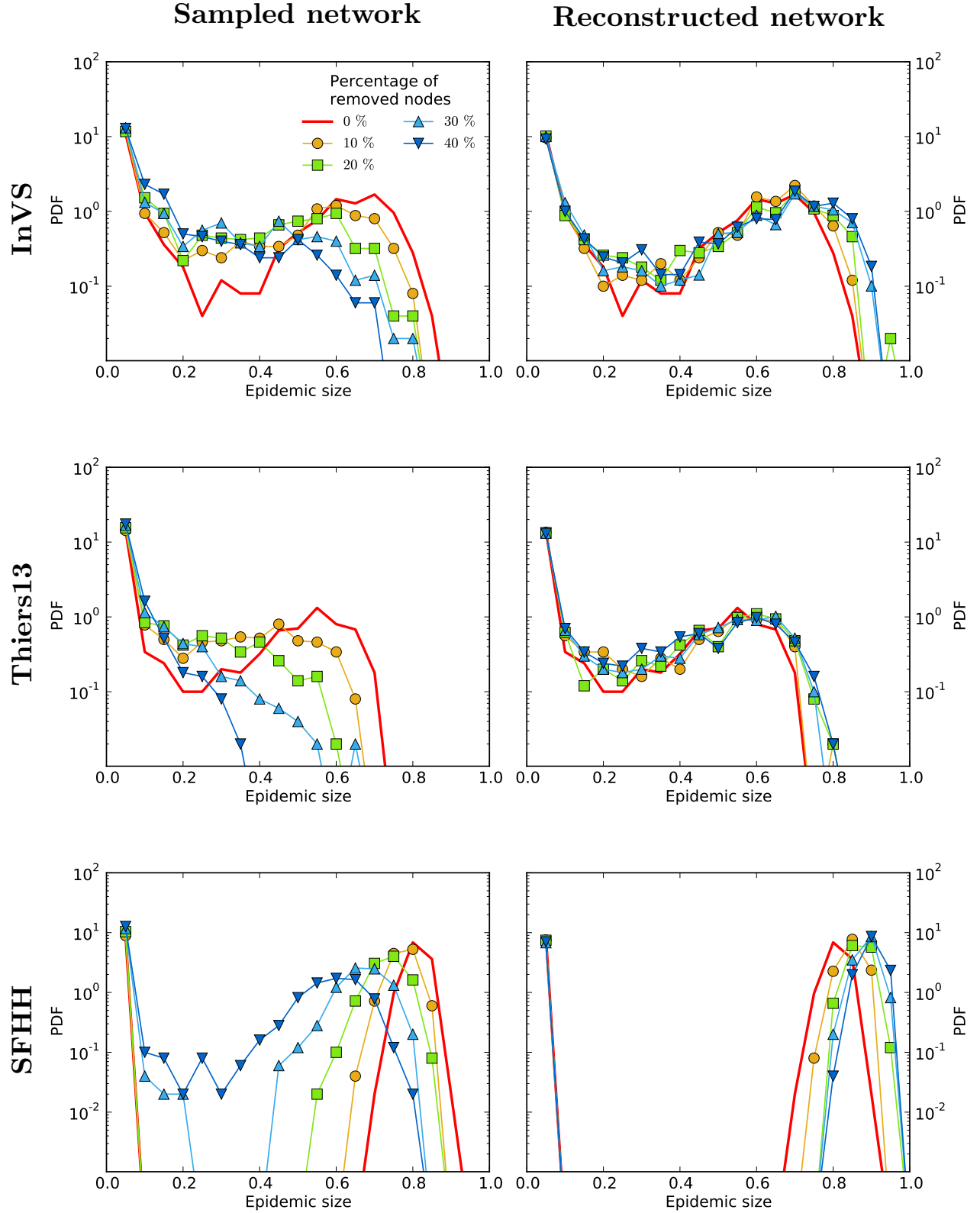


Supplementary Figure 28. **WST method with constrained transitivity. Comparison of the outcome of SIR epidemic simulations performed on resampled and reconstructed contact networks.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (*InVS*) or  $\beta/\mu = 100$  (*Thiers13* and *SFHH*). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.

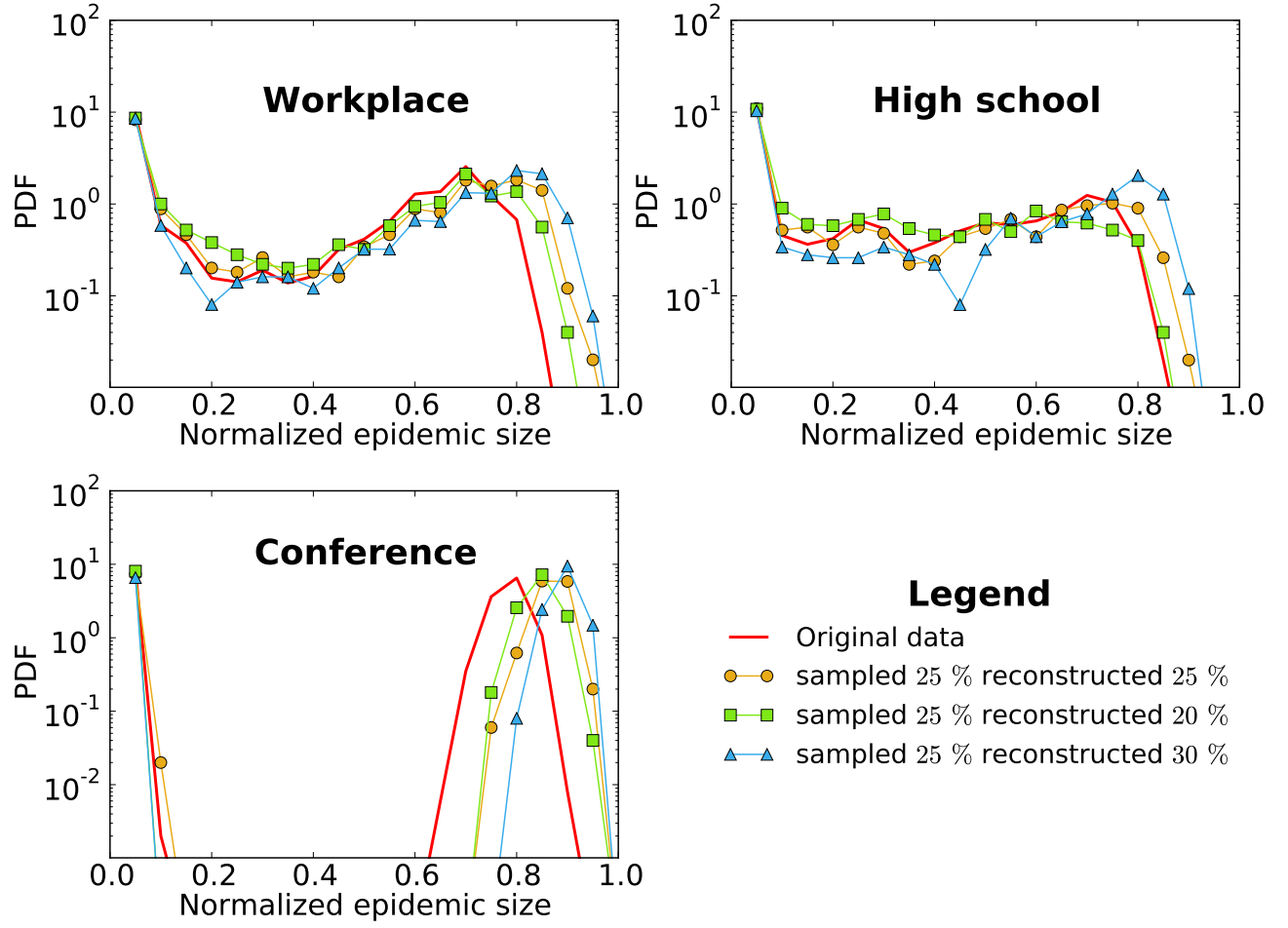


Supplementary Figure 29. **Method WST on link-shuffled network. Comparison of the outcome of SIR epidemic simulations performed on resampled and reconstructed contact networks.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (*InVS*) or  $\beta/\mu = 100$  (*Thiers13* and *SFHH*). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.





Supplementary Figure 30. **Method WST on time-shuffled network. Comparison of the outcome of SIR epidemic simulations performed on resampled and reconstructed contact networks.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (*InVS*) or  $\beta/\mu = 100$  (*Thiers13* and *SFHH*). The case  $f = 0$  corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.



Supplementary Figure 31. **Uncertainty on the sampling fraction. Comparison of the outcome of SIR epidemic simulations performed on contact networks where 25 % of nodes were removed, and reconstructed with different values of the assumed sampling fraction.** Distribution of epidemic sizes (fraction of recovered individuals) at the end of SIR processes simulated on top of either resampled (left column) or reconstructed (right) contact networks, for different values of the fraction  $f$  of nodes removed. The parameters of the SIR models are  $\beta = 0.0004$  and  $\beta/\mu = 1000$  (*InVS*) or  $\beta/\mu = 100$  (*Thiers13* and *SFHH*). The case “Original data” corresponds to simulations using the whole data set, i.e., the reference case. For each value of  $f$ , 1,000 independent simulations were performed.

## SUPPLEMENTARY NOTES

### Supplementary note 1: Effect of sampling on the temporal network of contacts

As described in the main text, we consider temporally resolved networks of contacts  $\mathcal{T}$  in a population  $\mathcal{P}$  of  $N$  individuals and we perform a resampling experiment by selecting a subpopulation  $\tilde{\mathcal{P}}$  of these individuals, of size  $\tilde{N} = (1 - f)N$ . We assume that only the contacts occurring among the subpopulation  $\tilde{\mathcal{P}}$  are known and we compare the properties of the corresponding resampled subnetwork  $\tilde{\mathcal{T}}$  with those of the original network.

Supplementary Figure 9 shows how population sampling affects several statistical properties of the contact networks. On the one hand, the degree distribution of the aggregated network of contacts systematically shifts towards smaller degree value. This is expected as each remaining node has in the resampled network a degree which is at most its degree in the original network, and is strictly smaller if some of its neighbours are not part of the resampled population. On the other hand, the statistical distributions of several quantities of interest are not affected by sampling: This is the case of the quantities attached either to single contacts or to single links, namely contact and inter-contact durations, number of contacts per link and link weights (the weight of a link is given by the total duration of the contacts between the two corresponding nodes).

Moreover, as shown in Supplementary Figure 10, the density of the aggregated network, i.e. the ratio between the number of links and the number of possible links, is on average conserved by the random resampling procedure. It varies however for different realisations of the resampling, and the corresponding variance increases with the fraction  $f$  of excluded nodes.

Supplementary Figure 11 shows how the average clustering coefficient of the aggregated network varies with the resampling: notably, it remains high and close to its original value until large values of  $f$  are reached. The transitivity of the network, defined as three times the number of triangles divided by the number of connected triplets (connected subgraphs of 3 nodes and 2 edges), is even less affected than the clustering coefficient by the resampling procedure.

In the case of structured populations, Supplementary Figures 12 & 13 show that the stability of the resampled network's density holds at the more detailed level of the contact matrices of link densities. In such matrices, the element  $(i, j)$  is given by the number of links between individuals of groups  $i$  and  $j$ , normalised by the total number of possible links between these two groups (if  $n_i$  denotes the number of individuals in group  $i$ , the number of possible links is equal to  $n_i n_j / 2$  for  $i \neq j$  and to  $n_i(n_i - 1) / 2$  for  $i = j$ ). These figures clearly illustrate how the diagonal and block-diagonal structures are preserved, and Supplementary Figure 10 gives a quantitative assessment of this stability by showing that the cosine similarity between contact matrices between the resampled and original aggregated contact networks remains high even for when a large fraction of the nodes are excluded.

We moreover illustrate in Supplementary Figures 14 and 15 the difference in statistical properties of contacts and links within and between groups, still for structured populations:

- the distributions of contact durations are indistinguishable;
- the distribution of link weights (aggregated contact durations) is broader for links between individuals belonging to the same group than for links joining individuals of different groups;
- this is due to the difference in the distributions of numbers of contacts per link, which is broader for links within groups than for links between groups;
- the distributions of inter-contact durations differ also slightly, with smaller averages for within-group links.

Most importantly, all these properties and distributions remain stable under resampling, showing that reliable information on the distributions of contact and inter-contact durations, aggregated contact durations, numbers of contacts per link, can be obtained in the resampled data, including the statistical differences between links joining members of different groups and links between two individuals of the same group.

### Supplementary note 2: Properties of the reconstructed contact networks

As described in the main text and in particular in the Methods section, we construct a surrogate set of contacts concerning the  $fN$  individuals excluded by the resampling. We compare here the properties of the resulting contact networks (obtained by merging the resampled contact network  $\tilde{\mathcal{T}}$  and the surrogate set of contacts) and of the original contact network,  $\mathcal{T}$ .

Supplementary Figure 16 shows that the degree distribution, which is not constrained by the reconstruction procedure, deviates from the original distribution. On the other hand, the distributions of contact durations, inter-contact

durations, number of contacts per link and link weights are preserved. Moreover, the link density contact matrices of the reconstructed networks (Supplementary Figure 17 & 18) share a high similarity with the original contact matrices, even for high fractions of nodes excluded (Supplementary Figure 19).

For completeness, we also compute the contact matrices in contact time density (CMT), in which each element  $(i, j)$  is given by the total time in contact between individuals of groups  $i$  and  $j$ , normalised by the total number of possible links between these two groups: it gives the average time spent in contact by two random individuals of groups  $i$  and  $j$ . Supplementary Figures 19, 20 and 21 show that the structure of these matrices is well recovered by the reconstruction methods, with high similarity with the original matrices.

### Supplementary note 3: Phase diagram of the SIS model for the conference and high school data sets

We observe for the high school and the conference the same effect on the phase diagram of the SIS model as in the workplace: sampling leads to a shift of the epidemic threshold to higher values and thus to an underestimation of the epidemic risk. The phase diagram and the epidemic threshold are estimated more accurately by using reconstructed networks, thus giving a better evaluation of the epidemic risk (Supplementary Figures 22 & 23).

### Supplementary note 4: Sensitivity analysis

In the main text, we have considered values of the SIR model parameters leading to a non-negligible epidemic risk and a value of  $\beta$  corresponding to slow processes. We consider here several other values of the parameters, corresponding either to faster processes (Supplementary Figures 24 - 26) or to smaller epidemic risk (Supplementary Figure 27). In all cases, simulations performed on the resampled contact networks lead to a strong underestimation of the epidemic sizes, with distributions shifting to smaller values as  $f$  increases, while the use of reconstructed data sets leads to a better estimation and generally speaking a slight overestimation of the epidemic risk.

## SUPPLEMENTARY METHODS

### Detailed alternative reconstruction methods

We give here details on the alternative reconstruction methods mentioned in the main text, which use less information than the **WST** method. In each case we consider the same setup as the complete method: a population  $\mathcal{P}$  of  $N$  individuals (the nodes of the contact network), potentially organised in groups, for which we know all the contacts taking place among a subpopulation  $\tilde{\mathcal{P}}$  of size  $\tilde{N} = (1 - f)N$ . For the remaining  $n = N - \tilde{N} = fN$  individuals, no contact information is available, but we know to which group they belong. We also have access to the overall activity timeline, *i.e.*, to the successive intervals during which contacts can happen (daytimes), and are excluded (nights and weekends). The alternative reconstruction methods are the following:

**0:** We perform the reconstruction using only the network density and the average link weight, both measured in the resampled network  $\tilde{\mathcal{T}}$ . The algorithm goes as follows:

1. we measure in the resampled data:
  - the density  $\rho$  of links in the time-aggregated network;
  - the average link weight  $\langle w \rangle_s$  (the weight of a link is defined as the total contact time between the two linked nodes);
2. we compute the number of links  $e$  that must be added to keep the network density constant when we add the  $n$  excluded nodes;
3. we construct  $e$  links according to the following procedure:
  - a node  $i$  is randomly chosen from the set  $\mathcal{P} \setminus \tilde{\mathcal{P}}$  of excluded nodes;
  - a node  $j$  is randomly chosen from the set  $\mathcal{P} \setminus \{i\}$  of all other nodes;
  - we compute  $n_{ij} = \langle w \rangle_s / \Delta t$ , where  $\Delta t = 20s$  is the temporal resolution of the data set, and we randomly choose  $n_{ij}$  time windows of length  $\Delta t$  within the activity windows defined by the activity timeline as contact events between  $i$  and  $j$ .

**W:** We perform the reconstruction using only the network density and the distribution of link weights, both measured in the resampled network  $\tilde{\mathcal{T}}$ . The algorithm goes as follows:

1. we measure in the resampled data:
  - the density  $\rho$  of links in the time-aggregated network;
  - the list  $\{w\}$  of link weights (the weight of a link is defined as the total contact time between the two linked nodes);
2. we compute the number of links  $e$  that must be added to keep the network density constant when we add the  $n$  excluded nodes;
3. we construct  $e$  links according to the following procedure:
  - a node  $i$  is randomly chosen from the set  $\mathcal{P} \setminus \tilde{\mathcal{P}}$  of excluded nodes;
  - a node  $j$  is randomly chosen from the set  $\mathcal{P} \setminus \{i\}$  of all other nodes;
  - from  $\{w\}$ , we draw the weight  $w_{ij}$  of the link  $ij$ ;
  - we compute  $n_{ij} = w_{ij}/\Delta t$ , where  $\Delta t = 20s$  is the temporal resolution of the data set, and we randomly choose  $n_{ij}$  time windows of length  $\Delta t$  within the activity windows defined by the activity timeline as contact events between  $i$  and  $j$ .

**WS:** We perform the reconstruction using the network density, the distributions of link weights for internal (within groups) and external (between groups) links, and the structure of the aggregated network given by the link density contact matrix, all measured in the resampled network  $\tilde{\mathcal{T}}$ . The algorithm goes as follows:

1. we measure in the resampled data:
  - the density  $\rho$  of links in the time-aggregated network;
  - a row-normalised contact matrix  $C$ , in which the element  $C_{AB}$  gives the probability for a node in group  $A$  to have a link to a node of group  $B$ ;
  - the lists  $\{w\}^{\text{int}}$  and  $\{w\}^{\text{ext}}$  of link weights for respectively internal and external links (internal links are links between nodes that belong to the same group, external links are links between nodes from different groups);
2. we compute the number of links  $e$  that must be added to keep the network density constant when we add the  $n$  excluded nodes;
3. we construct  $e$  links according to the following procedure:
  - a node  $i$  is randomly chosen from the set  $\mathcal{P} \setminus \tilde{\mathcal{P}}$  of excluded nodes;
  - knowing the group  $A$  that  $i$  belongs to, we extract at random a target group  $B$  with probability given by  $C_{AB}$ ;
  - we draw a target node  $j$  at random from  $B$  (if  $B = A$ , we check that  $j \neq i$ );
  - depending on whether nodes  $i$  and  $j$  belong to the same group or not, we draw from  $\{w\}^{\text{int}}$  or  $\{w\}^{\text{ext}}$  the weight  $w_{ij}$  of the link  $ij$ ;
  - as for the **W** method, we extract at random  $w_{ij}/\Delta t$  contact events of length  $\Delta t = 20s$  within the activity timeline.

**WT:** We perform the reconstruction using the network density, the distribution of link weights and the temporal structure of the contacts given by the distributions of contact durations, inter-contact durations, number of contacts per link and initial waiting times before the first contact, all measured in the resampled network  $\tilde{\mathcal{T}}$ . The algorithm goes as follows:

1. we compute from the activity timeline the time  $T_u$  as the total duration of the periods during which contacts can occur.
2. we measure in the resampled data:
  - the density  $\rho$  of links in the time-aggregated network;
  - the list  $\{\tau_c\}$  of contact durations;
  - the list  $\{\tau_{ic}\}$  of inter-contact durations;
  - the list  $\{p\}$  of numbers of contacts per link;
  - the list  $\{t_0\}$  of initial waiting times before the first contact for each link;

3. we compute the number of links  $e$  that must be added to keep the network density constant when we add the  $n$  excluded nodes;
4. we construct  $e$  links according to the following procedure:
  - (a) a node  $i$  is randomly chosen from the set  $\mathcal{P} \setminus \tilde{\mathcal{P}}$  of excluded nodes;
  - (b) a node  $j$  is randomly chosen from the set  $\mathcal{P} \setminus \{i\}$  of all other nodes;
  - (c) we draw from  $\{p\}$  the number of contact events  $p$  taking place over the link  $ij$ ;
  - (d) from  $\{t_0\}$ , we draw the initial waiting time  $t_0$  before the first contact;
  - (e) from  $\{\tau_c\}$ , we draw  $p$  contact durations  $\tau_c^k$ ,  $k = 1, \dots, p$ ;
  - (f) from  $\{\tau_{ic}\}$ , we draw  $p - 1$  inter-contact durations  $\tau_{ic}^m$ ,  $m = 1, \dots, p - 1$ ;
  - (g) while  $t_0 + \sum_k \tau_c^k + \sum_m \tau_{ic}^m > T_u$ , we repeat steps (c) to (f);
  - (h) from  $t_0$ , the  $\tau_c^k$  and  $\tau_{ic}^m$ , we build the contact timeline of the link  $ij$ ;
  - (i) finally, we insert in the contact timeline the breaks defined by the activity timeline.

### Reconstruction with fixed transitivity

In order to constrain the transitivity to its value measured in the resampled data, we add to the WST algorithm the following elements:

1. we measure in the resampled data the transitivity  $\sigma_0$  of the time-aggregated network;
2. for the construction of each link of a node  $i$ :
  - we calculate the current transitivity  $\sigma$  of the network;
  - we list the potential targets  $j$  in two lists  $C_\Delta$  and  $C_\wedge$ , depending on whether the creation of a link between  $i$  and  $j$  would close a triangle or not;
  - – if  $\sigma < \sigma_0$ , we draw a target node  $j$  at random from  $C_\Delta$  such that  $i$  and  $j$  are not linked;
  - else if  $\sigma > \sigma_0$ , we draw a target node  $j$  at random from  $C_\wedge$  such that  $i$  and  $j$  are not linked.

We show in Supplementary Figure 28 the outcome of simulations performed on reconstructed data sets using this modified algorithm.